Reconnaissance de scènes multimodale embarquée

David Blachon

Communauté
UNIVERSITÉ Grenoble Alpes

THÈSE

Pour obtenir le grade de

DOCTEUR DE LA COMMUNAUTÉ UNIVERSITÉ GRENOBLE ALPES

Spécialité : **Informatique**

Arrêté ministériel : 7 août 2006

Présentée par

David Blachon

Thèse dirigée par **Laurent Besacier**

préparée au sein du **Laboratoire d'Informatique de Grenoble (LIG)**
et de l'**Ecole Doctorale Mathématiques, Sciences et Technologies de l'Information, Informatique**

Reconnaissance de scènes multimodale embarquée

Thèse soutenue publiquement le **29 février 2016**,
devant le jury composé de :

Mme Catherine Garbay
Directrice de recherche CNRS, LIG, Grenoble, Présidente
M. Yacine Bellik
Maître de conférences, HDR, Université Paris-Sud, LIMSI, Orsay, Rapporteur
M. Dan Istrate
Maître de conférences, HDR, UTC, Compiègne, Rapporteur
M. Georges Linarès
Professeur des universités, Université d'Avignon, LIA, Avignon, Examinateur
M. Laurent Besacier
Professeur des universités, Université Grenoble Alpes, LIG, Grenoble, Directeur de thèse
M. François Portet
Maître de Conférences, Grenoble-INP, LIG, Co-Encadrant de thèse

Remerciements

En commençant ma thèse de doctorat, je souhaitais rendre ce projet le plus épanouissant possible. L'enrichissement professionnel était une évidence pour moi à ce moment. Sortant d'une école d'ingénieurs et de mon stage de fin d'études, j'étais convaincu qu'une expérience dans le monde de la recherche allait être l'occasion de faire des découvertes scientifiques, techniques, méthodologiques. Mais je souhaitais aussi faire de ce projet une expérience humaine, enrichissante sur le plan personnel. À l'heure de clore cette période, je suis satisfait des expériences vécues et des rencontres faites. Mes remerciements vont aux personnes qui ont participé à mon épanouissement professionnel et personnel au cours de cette thèse.

Je souhaite d'abord remercier mes encadrants de thèse : Laurent Besacier, François Portet et Stéphan Tassart. J'avais été averti que l'encadrement multiple pouvait être une source de problèmes à cause des divergences d'opinions et d'idées. Mais dans mon cas, un équilibre a rapidement été trouvé. Je remercie François pour son soutien permanent, son implication dans mon travail, son ouverture d'esprit et la qualité de ses réflexions, commentaires et remarques tout au long de la thèse, jusqu'à la soumission du manuscrit. Je remercie Stéphan pour son implication malgré la distance géographique grandissante au cours de la thèse, sa rigueur scientifique et méthodologique, la qualité de ses suggestions et son ouverture d'esprit nécessaire à sa délicate position d'encadrant "industriel", à l'interface entre les deux partenaires académique et industriel. Je remercie Laurent pour la qualité de ses réflexions, conseils, décisions et orientations donnés tout au long de la thèse, qui ont permis de rendre le chemin moins sinueux. Au-delà de leurs qualités professionnelles, j'ai apprécié les personnalités de chacun d'eux et ça a été un plaisir de travailler avec eux.

Mes remerciements vont aussi aux "doctorants" du laboratoire, terme que j'employais pour désigner le groupe de personnes (doctorants, stagiaires, "post-doc", du même laboratoire ou d'un autre) avec qui j'ai effectué de nombreuses sorties sur mon temps libre. J'ai eu la chance de commencer en même temps que plusieurs autres doctorants de l'équipe et, rapidement, une complicité s'est installée. À tous, merci pour les moments passés, que je n'oublierai pas : Marwen, Quang, Sarah, Fréd, Marion, Johann, Sohnoun, Issam, Oussama, Juline, Yuko, Lucie, Mahmoud, Émilie, Pedro, Uyanga, Pat, Mateusz, Nadia, Andrew, Kim, Élodie, Alexis, Belén, Raquel et ceux que j'aurais pu oublier au moment de l'écriture. Merci aussi aux personnes rencontrées lors de mon implication dans l'association du laboratoire Lig-Synergy : Mehdi, Nicolas, Mario, Lauren et Cécile.

Je souhaiterais également remercier les membres des différentes équipes auxquelles j'ai été affecté pendant ma thèse. Chez STMicroelectronics, j'ai eu la chance de côtoyer un petit groupe dédié à l'algorithmique (Michel, Bill, Florian, Stéphan et Séb), inclus au sein d'une plus grande équipe dédiée au traitement acoustique menée par Hervé, équipe qui a ensuite connu plusieurs évolutions. Au LIG, j'ai travaillé au sein du groupe GETALP. Je garde un très bon souvenir de cette équipe, aux membres chaleureux et sympathiques.

Bien sûr, mes remerciements vont à ma famille : mes parents, ma sœur et Raphaël. De manière concrète, je les remercie pour l'intérêt constant porté à mes travaux. Je suis aussi conscient de l'héritage culturel et intellectuel reçu au cours de mon éducation et qui a eu

une influence sur les nombreux choix effectués pendant la thèse. Pour cela, je remercie mes parents.

Enfin, je souhaite remercier Maryam. Malgré la période délicate de fin de thèse, elle a fait preuve de patience et de compréhension et sa présence à mes côtés m'a été précieuse, pour passer les étapes difficiles et apprécier les bons moments. Au-delà de la thèse, je la remercie de faire partie de mon existence.

Table des matières

Résumé

Contexte : Cette thèse se déroule dans les contextes de l'intelligence ambiante et de la reconnaissance de scène (sur mobile). Historiquement, le projet vient de l'entreprise ST-Ericsson. Il émane d'un besoin de développer et intégrer un "serveur de contexte" sur smartphone capable d'estimer et de fournir des informations de contexte pour les applications tierces qui le demandent. Un exemple d'utilisation consiste en une réunion de travail où le téléphone sonne ; grâce à la reconnaissance de la scène, le téléphone peut automatiquement réagir et adapter son comportement, par exemple en activant le mode vibreur pour ne pas déranger. Les principaux problèmes de la thèse sont les suivants : d'abord, proposer une définition de ce qu'est une scène et des exemples de scènes pertinents pour l'application industrielle ; ensuite, faire l'acquisition d'un corpus de données à exploiter par des approches d'apprentissage automatique ; enfin, proposer des solutions algorithmiques au problème de la reconnaissance de scène.

Collecte de données : Aucune des bases de données existantes ne remplit les critères fixés (longs enregistrements continus, composés de plusieurs sources de données synchronisées dont l'audio, avec des annotations pertinentes). Par conséquent, j'ai développé une application Android pour la collecte de données. L'application est appelée RecordMe et a été testé avec succès sur plus de 10 appareils. L'application a été utilisée pour 2 campagnes différentes, incluant la collecte de scènes. Cela se traduit par plus de 500 heures enregistrées par plus de 25 bénévoles, répartis principalement dans la région de Grenoble, mais aussi à l'étranger (Dublin, Singapour, Budapest). Pour faire face au problème de protection de la vie privée et de sécurité des données, des mesures ont été mises en place dans le protocole et l'application de collecte. Par exemple, le son n'est pas sauvegardé, mais des coefficients MFCCs sont enregistrés.

Définition de scène : L'étude des travaux existants liés à la tâche de reconnaissance de scène, et l'analyse des annotations fournies par les bénévoles lors de la collecte de données, ont permis de proposer une définition d'une scène. Elle est définie comme la généralisation d'une situation, composée d'un lieu et une action effectuée par une seule personne (le propriétaire du smartphone). Des exemples de scènes incluent les moyens de transport, la réunion de travail, ou le déplacement à pied dans la rue. La notion de composition permet de décrire la scène avec plusieurs types d'informations. Cependant, la définition est encore trop générique, et elle pourrait être complétée par des informations additionnelles, intégrée à la définition comme de nouveaux éléments de la composition.

Algorithmique : J'ai réalisé plusieurs expériences impliquant des techniques d'apprentissage automatique supervisées et non-supervisées. La partie supervisée consiste en de la classification. La méthode est commune : trouver des descripteurs des données pertinents grâce à l'utilisation d'une méthode de sélection d'attribut ; puis, entraîner et tester plusieurs classifieurs (arbres de décisions et forêt d'arbres décisionnels ; GMM ; HMM, et DNN). Également, j'ai proposé un système à 2 étages composé de classifieurs formés pour identifier les concepts intermédiaires et dont les prédictions sont fusionnées afin d'estimer la scène la plus probable. Les expérimentations non-supervisées visent à extraire des informations à

partir des données. Ainsi, j'ai appliqué un algorithme de regroupement hiérarchique ascendant, basé sur l'algorithme EM, sur les données d'accélération et acoustiques considérées séparément et ensemble. L'un des résultats est la distinction des données d'accélération en groupes basés sur la quantité d'agitation.

Abstract

Context : This PhD takes place in the contexts of Ambient Intelligence and (Mobile) Context/Scene Awareness. Historically, the project comes from the company ST-Ericsson. The project was depicted as a need to develop and embed a "context server" on the smartphone that would get and provide context information to applications that would require it. One use case was given for illustration : when someone is involved in a meeting and receives a call, then thanks to the understanding of the current scene (work meeting), the smartphone is able to automatically act and, in this case, switch to vibrate mode in order not to disturb the meeting. The main problems consist of i) proposing a definition of what is a scene and what examples of scenes would suit the use case, ii) acquiring a corpus of data to be exploited with machine learning based approaches, and iii) propose algorithmic solutions to the problem of scene recognition.

Data collection : After a review of existing databases, it appeared that none fitted the criteria I fixed (long continuous records, multi-sources synchronized records necessarily including audio, relevant labels). Hence, I developed an Android application for collecting data. The application is called RecordMe and has been successfully tested on 10+ devices, running Android 2.3 and 4.0 OS versions. It has been used for 3 different campaigns including the one for scenes. This results in 500+ hours recorded, 25+ volunteers were involved, mostly in Grenoble area but abroad also (Dublin, Singapore, Budapest). The application and the collection protocol both include features for protecting volunteers privacy : for instance, raw audio is not saved, instead MFCCs are saved ; sensitive strings (GPS coordinates, device ids) are hashed on the phone.

Scene definition : The study of existing works related to the task of scene recognition, along with the analysis of the annotations provided by the volunteers during the data collection, allowed me to propose a definition of a scene. It is defined as a generalisation of a situation, composed of a place and an action performed by a person (the smartphone owner). Examples of scenes include being in public transportation, being involved in a work meeting, walking in the street. The composition allows to get different kinds of information to provide on the current scene. However, the definition is still too generic, and it might be completed with additionnal information, integrated as new elements of the composition.

Algorithmics : I have performed experiments involving machine learning techniques, both supervised and unsupervised. The supervised one is about classification. The method is quite standard : find relevant descriptors of the data through the use of an attribute selection method. Then train and test several classifiers (in my case, there were J48 and Random Forest trees ; GMM ; HMM ; and DNN). Also, I have tried a 2-stage system composed of a first step of classifiers trained to identify intermediate concepts and whose predictions are merged in order to estimate the most likely scene. The unsupervised part of the work aimed at extracting information from the data, in an unsupervised way. For this purpose, I applied a bottom-up hierarchical clustering, based on the EM algorithm on acceleration and audio data, taken separately and together. One of the results is the distinction of acceleration into groups based on the level of agitation.

Table des figures

Liste des tableaux

CHAPITRE 1

Introduction

Motivation

Le travail de cette thèse se situe dans le domaine de l'informatique ambiante et ubiquitaire qui vise à proposer des services adaptés aux personnes en rendant les objets du quotidien intelligents. Des exemples concrets existent ou sont en cours de réalisation tels que les maisons intelligentes avec le contrôle de certains appareils électroniques pour assurer confort et sécurité ; les villes intelligentes dont l'éclairage peut s'allumer suivant la détection de présence ; les voitures intelligentes dont les premiers prototypes sont capables de se déplacer vers une destination avec une intervention du conducteur limitée voire nulle. Le point commun de ces exemples est la nécessité de percevoir son environnement.

Le *smartphone* (anglicisme pour le téléphone dit intelligent) est adapté à l'intelligence ambiante. En effet, depuis une dizaine d'années, nous assistons à une très forte expansion de la vente de ces appareils [1]. Les usages se sont multipliés grâce aux nombreuses applications disponibles et l'appareil est devenu tellement important pour certaines personnes qu'elles ne peuvent imaginer s'en passer durant la journée. En parallèle, les ressources de stockage, calcul et communication embarquées se sont aussi démultipliées pour permettre aujourd'hui la réalisation de tâches complexes telles que la navigation sur Internet ou le visionnage de vidéos. Pour ces raisons, le smartphone est un outil très adapté à l'objectif de l'intelligence ambiante.

Les services qui exploitent les ressources de l'appareil pour le rendre intelligent sont encore limités. Par exemple, la géo-localisation est utilisée pour adapter le contenu à une requête faite sur un moteur de recherche ; les accéléromètres sont employés pour déterminer l'orientation du téléphone et appliquer une rotation à l'affichage le cas échéant ; un capteur de luminosité ou de proximité est utilisé pour maintenir allumé ou éteindre l'écran automatiquement. Cette thèse est d'abord motivée par le besoin d'une perception plus fine du contexte. L'autre motivation est l'uniformisation de l'information et de la représentation de contexte qui ne nécessiterait plus des applications qu'elles emploient leurs propres représentations et qui faciliterait la communication des informations de contexte.

1. Alors qu'en 2007, il s'est vendu 122 millions d'appareils ; en 2014, le nombre de ventes sur l'année s'élève à 1,244 milliard. Voir les sites https://www.gartner.com/newsroom/id/910112 et https://www.gartner.com/newsroom/id/910112 pour les tableaux de valeurs

Le projet industriel

La thèse a été encadrée par une convention de type CIFRE [2] entre le Laboratoire d'Informatique de Grenoble et ST-Ericsson [3], puis STMicroelectronics. Le projet initial de la thèse résulte d'une nécessité de réponse de ST-Ericsson à l'industrie du smartphone. Le besoin est exprimé par le souhait de réaliser un *serveur de contexte*, embarqué sur le smartphone et dont les rôles principaux sont d'estimer le contexte ambiant de l'appareil à partir d'informations diverses et de l'exprimer pour le retourner aux applications tierces qui en font la demande. L'exemple donné pour illustrer l'objectif consiste à reconnaître la ***scène*** d'une réunion de travail. Dans cet exemple, l'application tierce est un module de gestion de la sonnerie du téléphone. Ce module peut tirer profit de l'information de contexte pour adapter le niveau de la sonnerie ou le type (sonnerie ou vibration) suivant la scène vécue et, ainsi, réagir convenablement à l'arrivée d'appels impromptus (par exemple, sonner si la sonnerie n'est pas dérangeante ou faire vibrer l'appareil dans le cas contraire).

L'information de contexte à délivrer n'est pas plus définie mais il est précisé qu'elle doit être exprimée sous différentes formes pour satisfaire aux requêtes des applications tierces. Par exemple, l'information peut être fournie suivant plusieurs niveaux d'abstraction : une application peut souhaiter connaître le nom de la scène ou seulement recevoir une information sur l'action effectuée dans celle-ci.

Le partenaire académique est l'équipe GETALP [4] du Laboratoire d'Informatique de Grenoble. Elle a été choisie suite à la volonté d'exploiter les données sonores pour la réalisation du système et par son expérience sur l'interaction vocale en contexte dans l'habitat intelligent.

Définition des problèmes

Le premier problème de la thèse réside dans la définition des scènes visées par le partenaire industriel. À partir de l'exemple de réunion fourni, nous extrapolons qu'il s'agit d'une situation de la vie courante et qu'elle est décrite à un niveau d'abstraction élevé. Le problème consiste à déterminer des exemples similaires à celui-ci, les regrouper et les abstraire pour le travail de recherche et, en particulier, l'évaluation du système. En outre, le problème consiste aussi à déterminer un modèle de scène suffisamment général pour décrire les exemples choisis et qui permette une description à différents niveaux de granularité à fournir aux applications tierces.

La méthode choisie pour la réalisation repose sur l'apprentissage automatique de classes par un système pour qu'il puisse ensuite les reconnaître. Cette méthode nécessite la constitution d'un corpus de données, annotées avec les concepts souhaités. Au préalable, il est nécessaire de déterminer les caractéristiques souhaitées pour les données. En particulier, pour

2. Conventions Industrielles de Formation par la Recherche.
3. Entreprise issue d'une *joint-venture* entre STMicroelectronics et Ericsson en 2009 et qui a cessé depuis.
4. Groupe d'Études pour la Traduction Automatique du Langage et de la Parole.

obtenir un corpus réaliste, les données doivent avoir été collectées dans un milieu naturel, donc dans les situations quotidiennes de volontaires. Il est possible de chercher un corpus existant ou d'effectuer une collecte dédiée. La collecte de données dans un milieu naturel soulève d'autres questions sur l'annotation ou la sécurité des données.

D'autres problèmes sont dus à la nature du smartphone, qui est l'appareil sur lequel le système de reconnaissance de scène doit être intégré. Le premier est lié aux ressources limitées de l'appareil (mémoire, batterie ou encore capacité de calcul) et à l'impact de l'usage des capteurs et du système sur ces ressources. Un autre problème porte sur les sources de données présentes sur l'appareil. Celles qui ont un intérêt potentiel se composent des capteurs de mesures physiques tels que le microphone, l'accéléromètre, le gyroscope, le magnétomètre, le baromètre, le capteur de luminosité ainsi que d'autres sources sur le fonctionnement du téléphone ou son usage telles que les applications utilisées, l'état de l'allumage de l'écran, le niveau de la batterie, la présence de communications sur un réseau sans fil (cellulaire, Wi-Fi, Bluetooth) ou la géo-localisation. Leur nombre est élevé et justifie la question de la pertinence ou de la combinaison des sources pour mieux tirer profit des informations collectées. Nous avons fait le choix de nous concentrer sur les capteurs physiques car ils réalisent des mesures directes du milieu ambiant, plus aptes, *a priori*, que les informations de fonctionnement de l'appareil, à représenter la scène. Ce choix est aussi cohérent avec l'activité du partenaire STMicroelectronics, qui consiste, entre autres, à fabriquer des puces matérielles.

En plus d'être nombreuses, la disponibilité des sources n'est pas assurée et peut varier d'un appareil à l'autre. On peut considérer que le microphone est présent sur tous les smartphones, mais ce n'est pas le cas des autres capteurs physiques. Par conséquent, cet élément doit être pris en compte dans l'évaluation du système ou dans l'adaptation de son comportement.

Enfin, on note le problème de la gestion de la vie privée. En effet, la collecte et l'usage des sources évoquées peut soulever des questions relatives aux traitements effectués. Par exemple, le microphone et le GPS sont des sources dont les données sont sensibles et qu'il faut traiter avec prudence.

Objectifs de la thèse

L'objectif du travail de thèse est de proposer des solutions pour la reconnaissance multimodale de scènes, embarquée sur un smartphone. Pour parvenir à ce résultat, et suivant les problèmes précédemment énoncés, nous définissons des objectifs.

1. Proposer une définition générale de scène qui exprime plusieurs niveaux ou éléments de description ; définir un ensemble d'exemples pour les expérimentations ;
2. Faire l'acquisition d'un corpus de données annotées suivant les exemples considérés ;
3. Proposer une ou plusieurs solutions au problème de reconnaissance de scène ; en particulier, les solutions doivent intégrer la possibilité de combiner les sources et offrir

plusieurs niveaux de description de la scène.

Plan du manuscrit

Pour aborder le concept de scène, le chapitre 2 sur l'état de l'art étudie le concept du contexte car son sens paraît proche de celui d'une scène et les travaux qui le considèrent sont plus nombreux. En outre, l'étude du contexte permet d'aborder les concepts de conscience du contexte et de représentation du contexte, pertinents pour la reconnaissance de scène. Ensuite, nous avons étudié des travaux qui appliquent la représentation du contexte à différentes tâches, en lien avec la reconnaissance de scènes. L'étude de ces travaux illustre les méthodes appliquées, problèmes rencontrés, sources de données employées, descripteurs calculés et classifieurs utilisés. Nous proposons une première comparaison des sources de données et une description des classifieurs jugés les plus pertinents ou populaires.

Le chapitre 3 aborde le problème de la collecte du corpus. Les caractéristiques souhaitées pour les données sont d'abord décrites. Après un état de l'art du domaine, nous décrivons le protocole et la réalisation d'une collecte de données dédiée. Plusieurs problèmes liés à l'annotation, à la sécurité des données et à la gestion de la vie privée sont abordés et des solutions sont proposées. À la fin du chapitre, deux collectes réalisées sont décrites.

Le chapitre 4 est la pierre angulaire du manuscrit. Il fait d'abord le bilan des résultats des deux premiers chapitres pour proposer une définition de scène qui généralise le concept et met en évidence les éléments de sa composition. Ensuite, les contraintes d'application industrielle, de collecte et de définition de la scène sont résumées pour introduire et justifier les solutions proposées au problème de reconnaissance de scène.

Le chapitre 5 présente les résultats d'une approche de classification supervisée de vecteurs de scène en réponse au problème de reconnaissance de scène. Des classifieurs et algorithmes de l'état de l'art sont employés, l'évaluation de la solution est effectuée suivant plusieurs configurations de sources de données et dans des conditions de cas d'utilisation réalistes. Les résultats rapportés constituent une référence de performances. Une seconde expérimentation est décrite qui aborde le problème de la reconnaissance de scène par la détection de transitions.

La solution du chapitre 5 est insuffisante car la réponse fournie par le classifieur ne permet pas plusieurs niveaux de description. C'est pourquoi dans le chapitre 6, nous proposons une solution alternative basée sur un système composé de plusieurs modules de reconnaissance combinés pour obtenir la description de la scène. En outre, nous présentons une expérimentation d'analyse non-supervisée des données collectées dans le but d'identifier des motifs et de les interpréter pour compléter la compréhension de la scène.

CHAPITRE 2

État de l'art

Le chapitre est organisé suivant quatre sections. Dans un premier temps, la notion de contexte est étudiée, de manière générale puis dans le cadre de l'informatique ambiante. À partir de cette définition, nous décrivons comment le contexte peut être représenté dans les systèmes, en présentant une méthode consensuelle. Le contexte est étudié comme une première approximation de la scène car les travaux sur le sujet sont plus abondants et les définitions plus nombreuses.

Par la suite, le chapitre présente des exemples concrets de travaux de reconnaissance de contexte. L'étude des travaux offre d'abord une vue d'ensemble des types de contexte que l'on peut reconnaître sur un smartphone ou, *a minima*, à partir de capteurs présents sur un smartphone. L'étude permet également d'identifier les méthodes, modèles et algorithmes sélectionnés pour ces travaux.

Les travaux étudiés permettent aussi d'identifier les sources de données pertinentes et, dans une troisième section, nous proposons une comparaison de celles-ci. Les critères portent sur l'analyse des travaux rapportés (notamment suivant la diversité des contextes et les performances notées) ainsi que sur des critères plus objectifs comme la consommation d'énergie, la disponibilité du capteur ou le caractère intrusif.

Enfin, dans une quatrième section, nous proposons une description des modèles et algorithmes pertinents rapportés dans l'étude des travaux existants.

2.1 Contexte et *conscience du contexte*

Par l'étude des définitions générales et appliquées du contexte, nous souhaitons retirer des éléments de description pertinents qui pourraient s'appliquer à la définition d'une scène. De plus, au-delà de la définition seule, c'est l'intégration du contexte dans les systèmes et la représentation de celui-ci qui nous intéresse pour pouvoir la transposer au concept de scène.

2.1.1 Définitions générales

De l'origine du mot à son application dans l'informatique ambiante, nous présentons dans cette section le contexte et les définitions proposées pertinentes pour notre étude.

2.1.1.1 Définition du contexte

L'origine du mot *contexte* remonte au moins au *XVI^e s.* selon le site du Centre National de Ressources Textuelles et Lexicales (CNRTL [1]). À l'origine, le mot *contexte* désigne "le texte d'un acte public ou sous seing privé", ainsi que "l'ensemble d'un texte par rapport à l'un de ses éléments, notamment dans la mesure où cet ensemble constitue une totalité signifiante et modifie ou affecte la valeur des éléments pris isolément". Le *contexte* désigne alors un texte qui donne du sens à un extrait de celui-ci pris isolément.

Puis l'usage du mot s'est élargi à d'autres domaines comme la linguistique et la musique. En linguistique, le *contexte* est décrit comme "l'ensemble des unités d'un niveau d'analyse déterminé [...] constituant l'entourage temporel (parole) ou spatial (écriture) d'une unité, d'un segment de discours dégagé par une analyse de même niveau". Le terme garde la notion d'ensemble qui entoure mais s'étend à d'autres domaines.

À partir du *XIX^e s.*, le terme désigne "[un] ensemble de circonstances liées, [une] situation où un phénomène apparaît, un événement se produit". L'usage de ce sens est devenu très fréquent vers 1960. Des syntagmes apparaissent alors comme le *contexte économique*, le *contexte culturel*, ou le *contexte géographique*, qui définissent un ensemble de circonstances dans un domaine et sont utilisés pour expliquer certains phénomènes.

Nous gardons cette définition comme point de départ de la définition des contextes dans ce sujet et nous allons la préciser dans le cadre de l'informatique ambiante.

2.1.1.2 Définition dans l'informatique ambiante

Dey et coll. (2001) ont analysé les définitions de contexte proposées par d'autres auteurs et ont remarqué qu'elles se classent en deux catégories : les définitions abstraites et pratiques. Ils ont proposé leur définition qui regroupe ses deux catégories et qui a ensuite été reprise par de nombreux auteurs (2010; 2001) :

> "*any information that can be used to characterize the situation of entities (i.e., whether a person, place, or object) that are considered relevant to the interaction between a user and an application, including the user and the application themselves. Context is typically the location, identity, and state of people, groups, and computational and physical objects.*"

De manière abstraite, le contexte en informatique représente toutes les informations utiles pour caractériser la situation d'entités ; et pertinentes pour décrire l'interaction d'un utilisateur et d'une application. On note également la proximité entre la notion de *contexte* et celle de *situation*. Cette définition est très générale mais n'est pas directement applicable dans un système.

À l'inverse, la deuxième partie de la définition énumère des éléments concrets du contexte : des entités telles que des personnes et des objets physiques ; des éléments de caractérisation de ces entités (localisation, identité et état). D'autres éléments sont proposés tels que l'heure, la saison ou la température (Brown et coll. (1997)), le temps et l'histo-

1. http://www.cnrtl.fr/etymologie/contexte visité le 09/04/2015 à 11h

rique (Chen et Kotz (2000)), l'état émotionnel, l'attention de la personne (Dey (1998)). Franklin et Flaschbart (1998) qualifient le contexte de *situation de l'utilisateur*, Ward et coll. (1997) d'*environnement de l'application*.

De la définition et des exemples présentés, nous retirons les éléments concrets de description du contexte que sont l'environnement, les personnes, les objets et le temps. Les autres éléments rapportés liés à la saison ou l'état émotionnel de la personne représentent un niveau de description plus précis et ne conviennent pas à la définition générale recherchée.

2.1.1.3 Les éléments du contexte

Dans la table 2.1, nous précisons les éléments de contexte mis en évidence dans la table. La localisation peut être géographique ou sémantique. Dans le premier cas, elle est décrite par un ensemble de coordonnées dans un système. Dans le second cas, on emploie un nom qui est associé à l'entité à caractériser (par exemple, on emploie l'Université Joseph Fourier pour désigner les locaux occupés par celle-ci). La localisation s'applique aux trois entités d'environnement, de personnes et d'objets énumérés.

Nous considérons que l'identité d'une entité représente des caractéristiques immuables telle que la fonction associée à un environnement. L'état correspond à des caractéristiques de l'entité qui peuvent être temporaires. Pour un environnement, on peut imaginer l'ambiance sonore ou lumineuse, la température ou l'humidité. L'état de personnes peut être décrit par leur nombre, leur posture, leur activité physique, les interactions sociales ou les émotions. L'état des objets peut décrire le fonctionnement. Pour un appareil comme le smartphone, on peut considérer le niveau de la batterie, la charge du processeur ou les ressources disponibles.

TABLE 2.1: Élements de contexte et leurs caractéristiques

	Localisation	**Identité**	**État**
Environnement	géographique ou sémantique	caractéristiques immuables	son, luminosité, température, humidité
Personnes			nombre, posture, activité physique, physiologique, émotion, interaction
Appareils			charge de batterie, processeur, ressources disponibles...
Temps			heure, date, historique

Les éléments fournis dans la table 2.1 paraissent applicables à la description d'une scène. Nous les considérons comme une base pour l'étude des travaux de reconnaissance de contexte, présenté dans la deuxième section du chapitre. Avant cela, nous abordons le problème de la représentation du contexte en informatique ambiante.

2.1.2 Mise en œuvre de la conscience du contexte

La capacité d'un système à se représenter son contexte a été qualifiée de conscience du contexte (*context awareness* en anglais), concept introduit par Schilit et Theimer (1994). Cette capacité permet à un système de "s'adapter en fonction de la position de l'utilisateur, des personnes et objets présents dans son environnement et de leurs changements au cours du temps" (1994). Nous décrivons une représentation composée de plusieurs couches d'abstraction successives qui transforment les mesures issues de capteurs en symboles. Nous décrivons aussi les traitements généraux associés à cette représentation.

2.1.2.1 Des symboles et transitions pour représenter le contexte

Historiquement, c'est dans le domaine de la vision par ordinateur que proviennent les premières propositions de modèles pour le contexte. Les modèles usaient de symboles et transitions pour représenter le contexte à l'image des scripts de Schank et Abelson (1977) ou des cadres (*frames* en anglais) de Minsky (1975).

Plus récemment, Crowley et coll. (2002) ont repris la notion de symboles et transitions pour décrire leur modèle du contexte. Selon leur définition, reprise par Coutaz et coll. (2005), le contexte peut être représenté au sein d'un graphe orienté qui représente l'univers des *contextes* possibles. Les nœuds du graphe sont les *contextes* et les arcs des transitions entre *contextes*. En outre, dans leur modèle, chaque *contexte* est un ensemble de *situations* qui sont autant d'instances particulières d'un contexte. Les *situations* se représentent également par les nœuds dans un graphe orienté dont les arcs modélisent les changements de *situation*. Ainsi, à tout moment, l'on se situe dans un état du graphe de *contextes* et également dans un état du graphe de *situations*. La figure 2.1 illustre ces deux graphes.

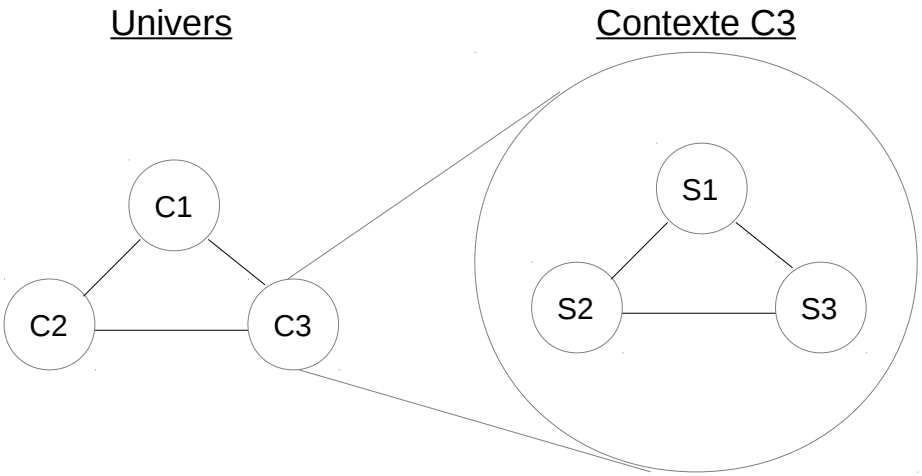


FIGURE 2.1: Représentation des contextes selon la définition de Crowley et coll. (2002)

Le modèle prévoit aussi la décomposition d'un contexte en entités, rôles et relations, décrits par des expressions de variables d'observations du système. Une *entité* est un regroupement de variables d'observations corrélées et peut être assimilée à un objet physique. Par

exemple, une personne est une entité. Elle peut être décrite par sa posture (représentée par une variable). Un ordinateur est un autre type d'entité, décrit par son état, représenté par une variable d'observation qui indique s'il est allumé ou éteint. Un *rôle* est une action qui peut être réalisée par une *entité*. Par exemple, le rôle d'une personne peut être de "travailler", celui de l'ordinateur est d'exécuter les commandes reçues. Enfin, une *relation* décrit l'interaction entre entités. Par exemple, lorsque la personne travaille avec son ordinateur, il y a une relation entre ces deux entités. Si la personne, se lève pour aller déjeuner, la rupture de la relation indique un changement de situation, voire un changement de contexte si la situation du déjeuner fait partie du contexte du travail ou d'un autre contexte.

La composition du modèle est intéressante car elle introduit une hiérarchie dans la représentation d'un contexte. De plus, la définition des entités, rôles et relations par des expressions de variables d'observations permet de relier le niveau d'abstraction des contextes d'une part, et le niveau des observations ou mesures d'autre part. Dans la suite, nous complétons la description du lien entre ces deux niveaux.

2.1.2.2 La perception du contexte

L'opération qui permet de transformer les observations en symboles du contexte est la *perception*. L'une des premières représentations de la perception du contexte est proposée dans les travaux de Hanson et coll. (1978). Dans leur article, les auteurs introduisent le concept de *pyramide du contexte*, dont une représentation est donnée dans la figure 2.2. Suivant le sens de lecture, la pyramide représente, de bas en haut, les niveaux d'abstraction des symboles extraits suite aux traitements consécutifs ou, de haut en bas, les hypothèses successives faites sur la composition de symboles abstraits en éléments de plus en plus concrets.

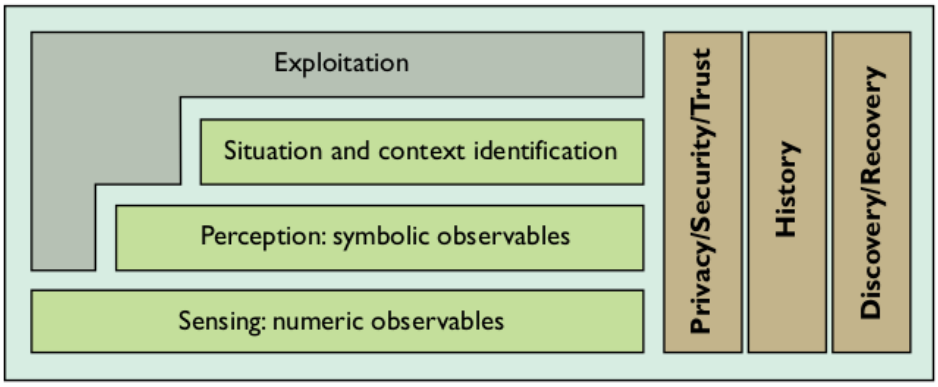


FIGURE 2.2: Représentation de l'architecture de perception pour les systèmes intelligents extraite de l'article de Coutaz et coll. (2005)

La pyramide de la figure 2.2 est composée de quatre niveaux mais trois sont réellement utilisés pour la représentation du contexte. La couche du bas est composée de capteurs qui fournissent des mesures, exprimées sous forme d'observations numériques. Au-dessus, la

couche intermédiaire représente l'étape de perception du système où les observations numériques sont transformées en variables symboliques. La couche du haut réalise l'identification du *contexte* ou de la *situation* (suivant les termes introduits avec le modèle de Crowley et coll. (2002)) à partir de ces observations. Une quatrième couche peut être superposée à la couche d'identification pour représenter l'exploitation de l'information de contexte. Cette couche est une interface entre les besoins exprimés par les applications et l'architecture d'identification de contexte.

2.1.2.3 Les traitements pour la conscience du contexte

Nous décrivons les traitements communs pour réaliser l'opération de perception du contexte. Pour cela, nous avons représenté dans la figure 2.3 la succession des étapes dans le cas d'un traitement hors-ligne. Les étapes du traitement en ligne peuvent être, ce que nous préciserons le cas échéant. Les blocs du schéma se situent après l'étape de mesure et avant l'étape d'identification du contexte.

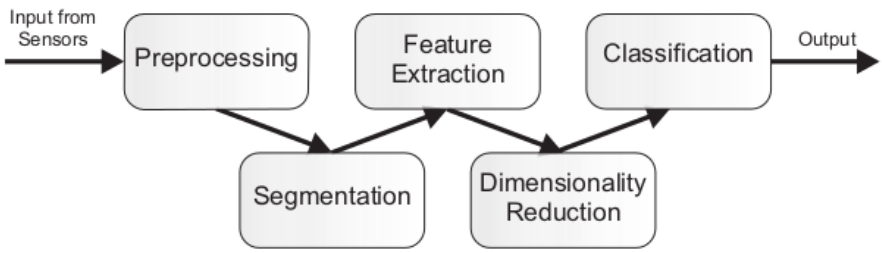


FIGURE 2.3: Traitement appliqué aux données sous forme de séries numériques temporelles, extrait de l'article de Avci et coll. (2010)

L'étape de segmentation consiste à découper une séquence de données en portions de taille plus petite et plus adaptée pour l'analyse (dans le cas du traitement en ligne, on parle de fenêtrage). La méthode la plus courante consiste à appliquer une fenêtre de taille fixée pour sélectionner les échantillons. La fenêtre est progressivement déplacée sur tout le signal.

L'étape de pré-traitement est une phase de mise en forme des données pour l'analyse ultérieure. Elle consiste souvent à filtrer le bruit présent avec les données. Diverses techniques peuvent être appliquées, suivant la connaissance *a priori* du signal. Parmi elles, nous citons le filtrage qui de bandes de fréquence sélectionnées et associées au bruit. Il est aussi possible de retirer des valeurs jugées aberrantes par leur amplitude, très différente de l'amplitude moyenne.

Lorsque le signal est segmenté et mis en forme, on applique des opérations pour retirer de l'information. Cela peut se faire avec une transformation vers un espace de représentation différent (telle que la transformée de Fourier). Il est également courant de calculer des descripteurs du signal. Ce sont des quantités censées représenter une partie de l'information transportée dans le signal. La moyenne ou la variance sont des exemples très fréquents de descripteurs statistiques. Les avantages de l'usage des descripteurs sont la réduction en

taille de la représentation des données et la concentration de l'information dans un nombre limité de quantités, ce qui les rend pertinents.

Il peut être difficile de déterminer les descripteurs qui sont réellement pertinents. C'est pourquoi, il est commun de procéder à une phase supplémentaire de réduction de dimension, dont le but est de réduire l'ensemble des descripteurs à un sous-ensemble optimal. Des méthodes de sélection d'attributs sont décrites dans le chapitre 5 de la thèse.

Enfin, la dernière étape consiste en l'estimation du contexte en fonction du sous-ensemble de descripteurs retenu. Brièvement, le processus de classification consiste à associer une étiquette issue d'un ensemble fini à un échantillon de données inconnu, à partir de la connaissance d'un corpus dont les données sont elles-mêmes associées à une étiquette de l'ensemble. L'étape préliminaire à la classification est appelée l'entraînement (ou l'apprentissage) et représente la construction de la connaissance des associations d'étiquettes aux données.

2.1.3 Bilan de la conscience du contexte en informatique ambiante

Nous retenons de cette section la description du contexte en éléments d'environnement, de personnes, d'appareils et de temps et la caractérisation de ces éléments par la localisation, l'identité et l'état. Cette description est pertinente pour l'application au concept de scène. De plus, les exemples d'états proposés sont repris dans la section suivante pour guider la revue des travaux de reconnaissance de contexte.

L'autre point important de la section est la représentation du contexte par la succession de trois niveaux : l'observation, qui fournit des mesures ; la perception, qui extrait de l'information des mesures pour les représenter sous forme de variables symboliques ; et l'identification du contexte qui manipule les symboles de contextes.

2.2 Revue de la reconnaissance d'éléments de contexte

Nous présentons un large éventail de travaux de reconnaissance de contexte qu'il est possible d'effectuer sur un smartphone. Les tâches de reconnaissance sont représentatives de certains éléments de contextes mis en évidence dans la section précédente. Ainsi, on retrouve la reconnaissance d'environnement (abordé par l'identification de la localisation et de l'ambiance sonore d'un lieu), la reconnaissance de l'état de personne (par l'activité physique et la situation sociale) et la reconnaissance du contexte du smartphone.

2.2.1 Reconnaissance de la localisation de l'utilisateur

La localisation d'une personne ou d'un objet est aujourd'hui possible en extérieur comme en intérieur grâce à des technologies et outils comme le GPS ou les radio-communications (antennes-relais, Wi-Fi, Bluetooth par exemple). Certaines de ces technologies, qui sont abordées plus en profondeur dans la section 2.3, sont présentes sur les smartphones et ont donné lieu à des travaux de reconnaissance de contexte.

Dans notre étude, nous relevons trois usages de la localisation. Le premier consiste à associer les coordonnées géographiques de la localisation à des labels textuels sémantiques. Il existe des cartes géographique enrichies qui réalisent de telles associations. C'est le cas d'OpenStreetMap (2014), un projet qui propose l'accès aux cartes et permet la participation à l'enrichissement par l'étiquetage de nouvelles zones géographiques. L'ensemble des labels couvre un vaste champ allant des infrastructures ou équipements aux éléments plus naturels comme un bois ou une prairie.

Le second usage de la localisation permet l'identification de points d'intérêt d'utilisateurs. Dans ce contexte, un point d'intérêt est un lieu de vie où l'on reste pendant une durée prolongée. Ravi et coll. (2008) rapportent l'évaluation d'un système qui analyse les traces de connexions du téléphone aux antennes-relais. La séquence des connexions est comparée à un historique de traces enregistrées, associées à certains événements comme les phases de recharge de la batterie du téléphone. Le système proposé tire profit de la comparaison pour prédire la prochaine occasion de recharge du téléphone suivant la séquence de traces courante. Les auteurs rapportent un taux d'erreur de la prédiction de 16 % à partir d'un corpus de traces de 9 mois, collecté par 80 volontaires. Ces résultats sont intéressants mais les auteurs précisent qu'ils ne sont envisageables que pour des utilisateurs avec un rythme de vie régulier. En outre, les auteurs évoquent le problème de la stabilité des traces. En effet, bien qu'immobile, un téléphone peut se connecter à différentes antennes en fonction des chemins suivis pas les ondes. Ces chemins dépendent de la topologie du lieu, en particulier de la présence de bâtiments dans les milieux urbains.

Azizyan et coll. (2009) se sont également intéressé à la reconnaissance de lieux par l'empreinte des bornes Wi-Fi accessibles. Un lieu est représenté par une empreinte, elle-même composée de la fraction de temps de visibilité de chacune des bornes accessibles dans le lieu. Par suite, une mesure de similarité est employée pour comparer deux empreintes. Évaluée sur un corpus d'empreintes provenant de 51 lieux et collecté par 4 volontaires, la méthode présente un taux de bonne classification de 70 %. Cependant, à l'image de la plupart des systèmes de localisation d'intérieur qui reposent sur les radio-fréquences (antennes-relais et Wi-Fi notamment), la précision de la localisation est limitée par la portée du signal : plus la portée est grande (quelques dizaine de mètres pour le Wi-Fi, plus encore pour les antennes-relais), plus la précision est faible.

Le troisième usage de la localisation consiste à détecter les entrées et sorties de bâtiments grâce au signal GPS. Marmasse et coll. (2000) ont proposé un système qui tire profit de l'opacité des bâtiments au signal GPS. Lors de l'entrée dans un bâtiment, le signal est rapidement perdu. Peu de temps après la sortie, le signal est retrouvé. En comparant la succession de coordonnées et d'indications temporelles, le système des auteurs peut identifier l'entrée et la sortie d'un bâtiment et, par extension, un point d'intérêt tel qu'il a été défini dans le paragraphe précédent. Les performances du système ne sont pas mentionnées, ni la précision temporelle ou géographique. Néanmoins, l'usage du GPS en milieu urbain donne lieu au problème connu des positions dites fantômes qui correspondent à de fausses mesures effectuées par l'appareil. Comme pour les signaux des antennes-relais, le chemin parcouru

par les ondes peut être influence par la topologie du lieu et des phénomènes de réflexion peuvent se produire et conduire à des erreurs de mesure.

Finalement, les performances relevées dans ces travaux sont encourageantes mais la précision de la localisation reste un problème important. De plus, certaines de ces sources ont une consommation élevée d'énergie (voir la section 2.3) ce qui les rend moins attrayantes.

2.2.2 Reconnaissance de l'environnement

Dans la section 2.1, nous avons caractérisé l'état d'un environnement par des exemples comme l'ambiance sonore ou lumineuse. Nous rapportons dans la suite plusieurs travaux qui ont exploité l'ambiance sonore des environnements en vue de les reconnaître. Nous ne considérons pas les travaux qui reposent sur l'usage d'images capturées par une caméra ou un appareil photo. En effet, dans l'usage quotidien, le smartphone passe beaucoup de temps dans une poche ou un sac, avec un champ de vision obstrué, ce qui limite l'usage de la caméra à des occasions ponctuelles. La source devient événementielle, or nous avons fait le choix de nous concentrer sur les sources continues.

La reconnaissance de l'environnement par son ambiance sonore repose sur l'hypothèse d'une homogénéité de celle-ci dans l'environnement. Le récent défi D-CASE[2] a proposé l'évaluation de systèmes de reconnaissance de scène sonore pour dix ambiances sonores quotidiennes telles que le bus, le restaurant, le bureau ou la station de métro. Auparavant, Peltonen et coll. (2002) ont été parmi les premiers à proposer un système de reconnaissance de scènes sonores. Les scènes considérées couvrent des environnements extérieurs (la rue, la "nature", le marché), des véhicules (voiture, bus, train), des lieux dits publics (restaurant, bar, supermarché), des lieux de travail (bureau, salle de classe ou de réunion, bibliothèque), des pièces du domicile et des lieux dits "réverbérants" comme les églises ou les gares. L'approche des auteurs considère des séquences sonores longues de 30 secondes et segmentées en portions plus courtes. Sur ces portions, des vecteurs de descripteurs acoustiques sont calculés. Les descripteurs choisis sont de différents types dans le but d'être évalués. Il y a des descripteurs temporels (nombre de passages par zéro du signal, énergie sur une fenêtre), des descripteurs spectraux (centroïde, bande passante et point *roll-off* qui caractérise la fréquence pour laquelle 95 % de l'énergie du spectre est représentée) et les coefficients cepstraux MFCC (*Mel Frequency Cepstral Coefficients* en anglais) qui sont calculés sur l'échelle Mel des fréquences. Des modèles composés entre autres de mélanges de distributions probabilistes gaussiennes (*Gaussian Mixture Models* en anglais, GMM) sont choisis pour représenter les ambiances et classifier les échantillons. Les auteurs rapportent des taux de reconnaissance d'à peine plus de 60 % pour 17 classes testées. Lorsqu'ils considèrent des problèmes de classification binaire, les auteurs indiquent des performances plus élevées : 94,7 % de bonne reconnaissance pour la classe des véhicules testée contre les autres classes, 86,3 % pour les lieux en intérieur testés contre les lieux en extérieur.

2. Proposé en 2013, le défi consiste en trois tâches de reconnaissance acoustique d'événements et de scènes à partir d'un corpus de sons fournis et d'un système de référence. La page Internet du défi est la suivante http://c4dm.eecs.qmul.ac.uk/sceneseventschallenge/.

D'autres études ont considéré des ambiances sonores similaires (Ma et coll. (2003)), ainsi que des ambiances dissociées d'un environnement, telle que la situation d'une conversation (Kern et coll. (2007)). Les méthodes sont également similaires : des descripteurs acoustiques sont calculés sur des fenêtres courtes (30 millisecondes pour Kern et coll. (2007)) ; des modèles génératifs sont employés pour modéliser les ambiances et les reconnaître. Ma et coll. (2003) et Kern et coll. (2007) rapportent l'usage de HMM pour la reconnaissance. Un HMM représente un processus stochastique défini par deux variables aléatoires, l'une qui produit des observations et l'autre qui gère les transitions entre états du HMM. Ma et coll. (2003) indiquent une précision de leur système de 91,5 % sur un corpus de 80 enregistrements de 3 secondes pour chaque scène, répartis à 75 % pour l'apprentissage et le reste pour le test. Kern et coll. (2007) indiquent avoir collecté un corpus long de 2 heures, dans 5 ambiances différentes et indiquent un taux de bonne reconnaissance de 88 %.

Une autre approche remarquée dans notre étude consiste à identifier des événements sonores particuliers afin d'en déduire l'environnement. Clarkson et coll. (1998) proposent une telle étude. Les sons considérés sont de différents types : parole, sonnerie d'un téléphone, bruit d'une voiture qui roule. Les scènes sonores étudiées sont la rue bruyante, le supermarché et le bureau. L'architecture du système est semblable à celles déjà mentionnées. Les échantillons sonores sont segmentés pour en calculer des descripteurs (des MFCC) et les événements sonores sont modélisés par des HMM. À partir des sons identifiés, les auteurs font l'hypothèse que chaque environnement est composé d'un ensemble de sons, plus ou moins différents d'un environnement à l'autre. Ainsi, les sons identifiés à un instant donné sont "comparés" aux compositions des différents environnements pour déterminer l'environnement le plus vraisemblable. Avec cette technique, les auteurs indiquent qu'il leur est également possible d'identifier les changements d'environnement par l'étude de l'évolution de la projection des sons.

Nous retenons de ces travaux les différents exemples de scènes de la vie quotidienne, basés sur des environnements ou des propriétés acoustiques particulières. Nous retenons également les approches employées qui reposent sur des vecteurs de descripteurs (souvent des MFCC) et des modèles (souvent des GMM ou HMM). Enfin, les performances évoquées pour ces approches sont encourageantes.

2.2.3 Reconnaissance d'activités physiques

Nous distinguons les activités suivant leur complexité sémantique. Les activités dites simples sont composées de mouvements fréquents, à l'image de la marche ou du déplacement à vélo. Les activités complexes sont les autres activités telles que l'action de cuisiner, de passer l'aspirateur ou de se laver les mains. Nous présentons d'abord les travaux de reconnaissance d'activités simples, puis ceux d'activités complexes.

2.2.3.1 La reconnaissance des activités simples et des postures

Les articles sur la reconnaissance d'activités physiques simples que nous avons étudiés considèrent un noyau commun d'activités simples qui sont la marche, la course, la montée et la descente d'escaliers, la bicyclette ainsi que des attitudes immobiles assise ou debout (Bao et Intille (2004), Kwapisz et coll. (2011), Shoaib et coll. (2014)). La source de données la plus populaire pour ces travaux est l'accéléromètre (Incel et coll. (2013)), mais nous présentons des études qui s'appuient sur d'autres sources comme le gyroscope ou le magnétomètre (Shoaib et coll. (2014)). La méthode employée pour la reconnaissance est commune aux différents articles étudiés : des descripteurs sont calculés sur des fenêtres longues de quelques secondes et un classifieur est entraîné pour la reconnaissance. Des descripteurs temporels sont employés tels que la moyenne, l'énergie, le nombre de passages par zéro et la corrélation suivant les données des axes de l'accéléromètre, ainsi que des descripteurs spectraux (entropie spectrale, premiers coefficients de la transformée de Fourier) (Bao et Intille (2004), Kwapisz et coll. (2011), Shoaib et coll. (2014)). Nous avons relevé de nombreux classifieurs employés : arbre de décision et réseau bayésien naïf (Bao et coll. (2004), Shoaib et coll. (2014)), perceptron (Kwapisz et coll. (2011)).

Les performances rapportées dans les différents articles sont très bonnes. Bao et Intille (2004) indiquent un taux de bonne classification global de 84 % et des taux encore plus élevés pour certaines activités telles que la marche (89,7 %), l'attitude immobile debout (95,7 %) ou la course (87,7 %) ; Kwapisz et coll. (2011) indiquent un taux global de 91,7 % ; Shoaib et coll. (2014) indiquent des taux de plus de 90 % dans différentes combinaisons de descripteurs et classifieurs. Les descripteurs et classifieurs mentionnés semblent donc constituer une bonne référence de méthode et de performances dans le domaine.

C'est pourquoi nous nous intéressons aux autres conditions expérimentales pour évaluer la pertinence des études relativement au contexte d'utilisation quotidienne du smartphone. D'abord, nous précisons les positions des capteurs considérées dans les études. En effet, l'accéléromètre et le gyroscope sont des sources inertielles et sont sensibles au mouvement. Par exemple, celui-ci peut être perçu différemment sur le bras ou à la taille. En outre, il est intéressant de considérer les positions réalistes pour un smartphone. L'étude de Bao et Intille (2004) a considéré des capteurs placés à la taille, au poignet, sur le bras, au genou et sur la cuisse. Parmi ces positions, la cuisse est celle qui conduit aux meilleurs résultats lorsqu'un seul capteur est considéré. La combinaison du capteur à la cuisse avec celui au poignet est le meilleur compromis entre le nombre de capteurs et la performance. L'étude de Kwapisz et coll. (2011) est effectuée avec un smartphone rangé dans la poche à l'avant du pantalon, ce qui se rapproche de la position de cuisse de Bao et Intille (2004). Shoaib et coll. (2014) ont considéré plusieurs positions possibles telles que les poches du pantalon, la ceinture, le poignet et le bras. Les bons résultats rapportés sont donc représentatifs de positions réalistes du smartphone.

Pour compléter l'évaluation des performances, nous considérons la composition du corpus et son exploitation dans les expérimentations. Nous avons déjà vu que les études portent

sur plusieurs activités physiques. En outre, Bao et Intille (2004) indiquent que 20 volontaires ont participé à la collecte et que les expérimentations ont été validées suivant deux protocoles. Le premier consiste à entraîner un classifieur avec les données d'un volontaire et à le tester sur d'autres données de ce même volontaire. Le second consiste à entraîner le classifieur sur les données de tous les volontaires sauf un, laissé de côté pour l'évaluation du classifieur (cette méthode est appelée en anglais *Leave One Subject Out Cross-Validation*, abrégée par LOSOCV). Les deux évaluations sont représentatives de deux cas d'utilisation différents. Le premier protocole est intéressant pour l'évaluation d'un classifieur partiellement ou totalement entraîné avec des données d'un utilisateur. Le second permet notamment de soustraire le potentiel biais d'apprentissage des données d'un utilisateur en soumettant le classifieur aux données d'un volontaire inconnu. Dans les deux cas, les protocoles sont réalistes pour notre application sur smartphone.

Kwapisz et coll. (2011) ont collecté des données de 30 volontaires et les ont exploitées suivant une validation croisée à 10 sous-ensembles. Contrairement à la LOSOCV précédemment mentionnée, cette méthode découpe le corpus entier aléatoirement en 10 sous-ensembles. Neuf sont employés pour l'entraînement et le dixième sert à l'évaluation. Le processus est répété afin que chaque sous-ensemble serve une fois au test. Shoaib et coll. (2014) ont appliqué la même évaluation. L'évaluation est représentative de l'utilisation d'un modèle entraîné avec les données de différents utilisateurs, dont celui qui sert au test.

Enfin, nous attirons l'attention sur la diversité des sources et leur combinaison. Bien que l'accéléromètre soit la source la plus populaire, son usage est limité, en particulier à cause de la dépendance au positionnement. Bao et Intille (2004) ont montré que l'usage de deux accéléromètres soigneusement placés améliore fortement l'usage d'un seul. Shoaib et coll. (2014) ont étudié les performances de classification de l'accéléromètre, du gyroscope et du magnétomètre, seul et combinés entre eux. Alors que le magnétomètre ne semble pas pertinent pour la tâche de reconnaissance d'activités physiques, le gyroscope semble compléter l'accéléromètre et la combinaison des deux permet d'atteindre des performances supérieures à celles obtenues lors de l'usage unique des capteurs. Ce dernier point est intéressant car dans la thèse, nous considérons l'usage d'un smartphone seul. Cela revient à considérer un seul accéléromètre sur le corps du volontaire. Mais le smartphone dispose d'autres sources de données, dont un gyroscope, pour compléter les données et améliorer l'inférence. Les conclusions de ces études sont donc aussi pertinentes quant à l'intérêt de disposer de plusieurs sources pour la tâche de reconnaissance d'activités physiques sur le smartphone.

2.2.3.2 La reconnaissance d'activités complexes

L'étude des activités physiques complexes semble pertinente pour la reconnaissance de scènes car ces activités représentent des situations avec un niveau de sémantique et d'abstraction élevé. Comme pour les activités simples, les études rapportées emploient essentiellement l'accéléromètre (Bao et Intille (2004), Dernbach et coll. (2012), Yan et coll. (2012)). Nous présentons d'abord une approche que nous qualifions de directe car, comme pour les activités simples, les classifieurs traitent des descripteurs calculés sur les données des cap-

teurs et reconnaissent les étiquettes des activités. Les activités considérées sont représentatives de la vie quotidienne avec, par exemple, passer l'aspirateur ou frotter une surface, déjeuner ou préparer le repas et des activités immobiles comme la lecture, le visionnage de la télévision ou le travail sur ordinateur. Les descripteurs et classifieurs employés sont similaires à ceux déjà mentionnés ; Dernbach et coll. (2012) ont tiré profit de l'outil Weka pour essayer des classifieurs supplémentaires comme un réseau bayésien (introduit dans la section 2.4.2.1) ou une table de décision (représentation de règles logiques). L'étude de Dernbach et coll. (2012) rapporte un taux de reconnaissance pour les activités complexes inférieur à 50 %, quel que soit le classifieur employé et en appliquant une méthode de validation croisée à 10 sous-ensembles. Les auteurs ont également collecté des activités simples et évalué les classifieurs sur ces activités. Les taux de reconnaissance relevés atteignent 90 % pour celles-ci.

Bao et Intille (2004) ont également effectué la collecte et l'évaluation d'activités complexes (dont le protocole a déjà été décrit précédemment). Les résultats rapportés pour les activités complexes sont globalement meilleurs que ceux de Dernbach et coll. : 77,3 % de reconnaissance pour l'activité de visionnage de la télévision, 88,7 % pour le déjeuner, 91,8 % pour l'activité de lecteur, 96,4 % pour le passage de l'aspirateur. Ces valeurs sont issues de la validation croisée avec un sujet laissé de côté, qui représente *a priori* une évaluation plus stricte que la validation croisée à 10 sous-ensembles employée par Dernbach et coll. (2012). Cependant, plusieurs accéléromètres sont portés par les volontaires de l'étude de Bao et Intille (2004). Les auteurs estiment à ce sujet que les accéléromètres positionnés sur le bras sont déterminants pour la reconnaissance des activités qui impliquent des mouvements du haut du corps ou des postures particulières (comme la lecture ou le travail sur ordinateur). À l'inverse, les volontaires de l'étude de Dernbach et coll. (2012) ont collecté les données avec un smartphone, positionné où ils le souhaitaient, laquelle position n'a pas été prise en compte dans l'article. Les conditions d'expérimentation de l'étude de Dernbach et coll. (2012) sont comparables à l'application visée par la thèse puisqu'un seul téléphone est employé, dans des positions qui peuvent varier.

Nous complétons l'étude de la reconnaissance d'activités physiques complexes avec l'article de Yan et coll. (2012) qui propose une approche compositionnelle qui consiste à représenter une activité complexe en une séquence d'activités simples. Le système proposé dans l'article est composé de deux étages d'inférence. Le premier étage applique une classification de vecteurs de descripteurs (calculés sur les données d'accéléromètre) pour reconnaître des activités simples. Les activités simples reprennent celles que nous avons déjà présentées (la marche, le déplacement dans des escaliers, les attitudes immobiles debout et assise) avec des nuances censées représenter des mouvements plus naturels telles que la posture assise avec les jambes qui bougent ou qui sont étirées ; la déambulation dans une pièce avec des arrêts occasionnels ou des hésitations de directions. L'évaluation de ce premier étage est effectuée sur des données de 5 volontaires, suivant une validation croisée à 10 sous-ensembles, avec des descripteurs temporels et spectraux déjà introduits et plusieurs classifieurs de l'outil Weka tels que l'arbre de décision, le réseau bayésien et réseau bayésien naïf. Les résultats

indiquent un taux de classification au-dessus de 89 %.

Pour le second étage, les auteurs ont essayé plusieurs stratégies de représentation des activités complexes par des activités simples. Brièvement, la séquence d'activités simples est d'abord représentée sous forme d'un vecteur, dont les valeurs indiquent le nombre d'instances de chaque activité simple unique. Également, les auteurs détectent la présence de motifs connus, sous la forme de sous-séquences d'activités simples, dont la présence ou l'absence seront intégrées dans le vecteur de représentation. Pour l'évaluation, le vecteur est comparé à une base de vecteurs annotés pour déterminer l'activité la plus probable. Les auteurs ont également entraîné un classifieur à reconnaître les activités complexes, directement à partir des vecteurs de descripteurs de l'accéléromètre. L'évaluation porte sur plusieurs activités complexes de travail, détente, cuisine, réunion, situées sur le lieu de travail ou au domicile. Au total, les auteurs indiquent 1102 instances d'activités complexes collectés par les 5 volontaires. Les taux de reconnaissance rapportés sont en faveur du système à deux étages qui présente une moyenne de 77 % sur les 1102 instances contre seulement 60 % pour le classifieur "direct". Le résultat est très intéressant car l'approche compositionnelle offre une autre ouverture (après la combinaison des différentes sources) pour aborder le problème du smartphone unique, au positionnement libre.

2.2.3.3 La reconnaissance de transports

Parmi les activités complexes, nous évoquons également les déplacements en transport. La première étude rapportée est celle de Sohn et coll. (2006) dont la démarche repose sur l'analyse des connexions du téléphone aux antennes-relais et la comparaison aux différents types de déplacement. Les connexions aux antennes sont représentées par une "empreinte", qui est un vecteur dont les valeurs indiquent la puissance du signal reçu par une antenne. En milieu urbain, un téléphone se situe dans la zone de couverture de plusieurs antennes-relais, ce qui justifie la présence de plusieurs valeurs. Pour la comparaison aux types de déplacement, les auteurs considèrent que plus le déplacement est rapide, plus les empreintes entre deux instants successifs sont différentes, puisque le déplacement éloigne de certaines antennes et rapproche d'autres, voire permet d'entrer dans la zone de couverture de nouvelles. À partir de cette observation, les auteurs ont comparé l'évolution des empreintes des antennes téléphoniques suivant les vitesses de déplacement, elles-mêmes associées à des modes de transports. La proximité de deux empreintes d'antennes est mesurée par une distance sur les vecteurs qui les représentent. La comparaison des distances entre antennes avec les modes de transport a permis aux auteurs de déterminer des intervalles de distances qui caractérisent les modes de transport. Ainsi, une distance inférieure à 5 dans l'espace des vecteurs d'empreintes d'antennes correspond à l'immobilité ; une distance comprise entre 5 et 10 correspond à la marche ; une distance supérieure à 10 indique un déplacement rapide et *a priori* motorisé. Les auteurs proposent un système entraîné à associer une étiquette de déplacement (parmi les trois catégories précédentes) aux distances entre les empreintes successives. Le système a été évalué et les auteurs rapportent un taux de classification de 85 %.

Le système de Reddy et coll. (2010) ressemble à celui de Sohn et coll. (2006). Les déplacements visés comportent la marche, la course, le vélo, le véhicule motorisé et l'absence de mouvement (l'immobilité). Comme Sohn et coll. avec les antennes-relais, Reddy et coll. tirent profit de la localisation de la personne *via* le GPS et la combine avec les données de l'accéléromètre. Les descripteurs employés se composent de la vitesse fournie par le GPS et, pour l'accéléromètre, de la variance et des composantes de la transformée de Fourier pour les fréquences de 1, 2 et 3 Hz. Par ailleurs, les auteurs comparent plusieurs classifieurs avec un système à deux étages composé d'un classifieur d'observations individuelles en entrée complété par un HMM dont le but est de lisser les prédictions du classifieur d'entrée et ainsi réduire les changements abrupts de classe. Leur base de données est composée d'échantillons collectés par 16 volontaires ayant passé approximativement 15 minutes dans les 5 moyens de déplacement considérés. L'arbre de décision s'est révélé le meilleur classifieur individuel avec une performance globale de 91 %. Lorsqu'il est combiné avec le HMM pour former le classifieur à deux étages, la performance atteint presque 94 %.

2.2.3.4 Bilan de la reconnaissance d'activités physiques

Les travaux de reconnaissance d'activités physiques simples indiquent de bons résultats et la méthode est éprouvée, ce qui est encourageant pour l'application au smartphone. À l'inverse, la reconnaissance d'activités complexes présente de moins bons résultats. De plus, les capteurs inertiels (accéléromètre et gyroscope) semblent les plus performants pour les mouvements simples. Cependant, ils sont très sensibles à la position sur les membres, ce qui affecte fortement les performances. Une solution observée plusieurs fois dans les études consiste à augmenter et diversifier les sources. Une autre solution consiste à composer le système avec des représentations intermédiaires.

2.2.4 Reconnaissance du contexte physique du smartphone

Le problème du contexte du smartphone a été évoqué dans la section précédente par la position variable de l'appareil sur le corps du volontaire qui le porte. Dans la suite, nous présentons plusieurs études pour compléter la description du contexte du smartphone.

2.2.4.1 Le problème du contexte du smartphone

Selon Lane et coll. (2010), le problème du contexte du smartphone réside dans la variabilité de son environnement immédiat qui peut avoir une influence sur les mesures des capteurs. Les auteurs proposent l'exemple d'un enregistrement sonore effectué dans une poche ou dans un sac et qui est perturbé par l'atténuation ou les frottements de l'appareil avec le tissu ou les objets à proximité. La position du smartphone sur le corps peut aussi changer la perception du mouvement, comme nous l'avons déjà introduit. Nous évoquons à ce sujet les travaux de Bouten et coll. (1997) qui ont montré que les accélérations mesurées sur le corps à l'occasion d'activités physiques varient en fréquence et en amplitude suivant le membre du corps où la mesure est effectuée. Hoseini-Tabatabaie et coll. (2013) confirment l'influence

de la position de l'appareil et évoquent également son orientation, qui peut modifier les mesures des capteurs inertiels. En effet, comme nous l'expliquons dans la section 6.1.1, la mesure fournie par les accéléromètres inclut une projection de l'accélération de pesanteur terrestre (ou gravité). Cette projection varie suivant l'orientation de l'appareil (et des accéléromètres embarqués) si bien que, pour une même activité et une même position de l'appareil sur le volontaire, les valeurs des mesures seront différentes suivant l'orientation.

Plusieurs solutions sont proposées pour résoudre le problème du contexte du smartphone. La première concerne la position de l'appareil et consiste à intégrer dans le corpus d'entraînement des données collectées par le smartphone positionné à différents endroits. C'est ce qu'ont proposé Lester et coll. (2006) pour leur système de reconnaissance d'activités physiques. Dans l'étude, trois positions sont considérées : l'épaule, le poignet et la taille. Les résultats du système entraîné et testé avec des données des trois positions sont aussi bons que les résultats de classifieurs entraînés et testés seulement sur les données d'une position.

Pour résoudre le problème de l'orientation du téléphone, une solution simple consiste à considérer des descripteurs invariants tels que la norme des valeurs des trois axes pour les capteurs inertiels (Hoseini-Tabatabaie et coll. (2013), Siirtola et coll. (2013)). Cette solution a le désavantage de perdre une partie de l'information contenue dans les valeurs des différents axes. Lu et coll. (2010) proposent une méthode qui transforme le vecteur d'accélération du repère de l'appareil dans le repère de coordonnées de l'utilisateur (ou repère terrestre). Sans boussole, il n'est pas possible de calculer les projections sur les 3 axes du repère, mais les auteurs parviennent à obtenir une projection sur l'axe vertical et une autre dans le plan horizontal, qui permet de connaître la direction du vecteur d'accélération.

2.2.4.2 Reconnaissance du contexte du smartphone

Une autre approche pour le problème du contexte du smartphone consiste à reconnaître la position de l'appareil. Cette solution est intéressante dans le cadre de cette thèse car l'information de position peut être intégrée aux informations de contexte fournies aux applications tierces.

Les études que nous rapportons emploient le microphone et l'accéléromètre pour reconnaître la position de l'appareil, par apprentissage sur des données collectées et annotées. Ainsi, Miluzzo et coll. (2010) proposent un système pour la classification d'ambiances sonores collectées et étiquetées dans deux positions : à l'intérieur et à l'extérieur d'une poche. La représentation des ambiances combine des MFCC et des modèles GMM et SVM (*Support Vector Machine* en anglais, système à vastes marges en français). Brièvement, le SVM représente les données dans un espace à plusieurs dimensions et détermine des hyperplans pour séparer l'espace en sous-espaces les plus homogènes possibles relativement aux classes apprises. Le corpus de données contient 15 minutes d'enregistrements sonores dans chacune des positions, affectées équitablement à l'entraînement et à l'évaluation. Les auteurs rapportent des taux de reconnaissance de 80 % pour les deux positions du smartphone.

Diaconita et coll. (2014) ont proposé un système de reconnaissance de position du smartphone à partir de la classification de sons. Le système émet un son aux caractéristiques

acoustiques connues ; le son est atténué par l'environnement immédiat de l'appareil (une poche de pantalon ou un sac par exemple) et enregistré par le système ; la classification s'opère sur l'enregistrement du son atténue pour identifier la position. Ainsi, les auteurs ont constitué un corpus d'enregistrements sonores dans plusieurs positions communes du smartphone (dans un sac à dos, posé sur une surface, tenu à la main ou rangé dans une poche) et dans plusieurs environnements sonores (le bus, le bureau, le tramway, l'extérieur et le train). Pendant les enregistrements, un signal particulier est joué et enregistré. Le volume du signal est fixé à la moitié du volume maximal possible sur le téléphone employé ce qui, d'après les auteurs, atténue fortement le bruit en environnement bruité ainsi qu'en environnement calme si l'appareil est dans un sac. L'évaluation vise à classifier l'environnement et la position du smartphone parmi toutes les combinaisons possibles des deux. Les auteurs ont comparé plusieurs classifieurs (arbre de décision, forêt d'arbres décisionnels (RF) ou encore GMM) et descripteurs (dont les MFCC qui ont donné les meilleurs résultats). Avec la combinaison de la forêt d'arbres décisionnels et des MFCC, les auteurs rapportent un taux de classification de 96 % suivant une évaluation en validation croisée à 10 sous-ensembles sur les enregistrements sonores. Une autre expérience de classification a été menée sans l'émission d'un son connu. Cette fois, la classification est effectuée à partir du seul son de l'environnement et vise à reconnaître la position et l'environnement. Le taux de classification obtenu chute à 71 % et les auteurs remarquent des confusions de position dans un même environnement. L'étude est intéressante car elle montre que, dans des conditions d'expérimentation et d'évaluation contrôlées, il est possible de reconnaître la position d'un appareil grâce au son.

Park et coll. (2012) ont abordé le problème par la reconnaissance des données d'accélérations. Celles-ci sont collectées auprès de 14 volontaires pendant une activité de marche réalisée avec le téléphone situé dans quatre positions considérées comme fréquentes : tenu à la main ; porté à l'oreille ; rangé dans la poche d'un pantalon ; rangé dans un sac. Un SVM est employé comme classifieur pour apprendre et reconnaître les positions. Les auteurs rapportent un taux global de 95 % de bonne classification, suivant une validation croisée avec un sujet laissé de côté (LOSOCV).

Enfin, nous rapportons les résultats d'une expérimentation de reconnaissance de position du smartphone que nous avons réalisée (Blachon et coll. (2014a)). Le corpus et le protocole sont décrits dans la section 3.4.2 et l'expérimentation est présentée plus précisément dans la section 6.2.1. Brièvement, nous avons fait appel à 19 volontaires qui ont réalisé un ensemble d'activités physiques. Pour la collecte, plusieurs smartphones étaient positionnés sur les volontaires (dans la poche avant du pantalon, dans le sac et à la main) et ont enregistré en continu des données sonores et d'accélérations. Avec des classifieurs à base d'arbres de décision (C4.5 et forêt d'arbres décisionnels), et suivant une validation croisée à 10 sous-ensembles, nous avons obtenu des résultats de classification supérieurs à 80 %, pour les trois configurations de descripteurs testés (accélérations et audio testés individuellement, puis en combinaison).

2.2.4.3 Bilan de la reconnaissance du contexte du smartphone

Avec cette étude, nous remarquons que le contexte du smartphone peut influencer la mesure et l'interprétation des données. Il paraît donc pertinent de le considérer dans le problème général de reconnaissance de scène. À notre connaissance, le contexte du smartphone est représenté par sa position et son orientation. Nous avons décrit plusieurs stratégies pour limiter l'influence de l'orientation. Nous avons aussi rapporté des travaux sur la reconnaissance du smartphone, qui peut être abordée avec des capteurs inertiels et acoustiques. Ces travaux sont pertinents pour l'application industrielle car l'information de la position du smartphone peut être intéressante à fournir.

2.2.5 Reconnaissance de situations sociales

La reconnaissance de situations sociales est un domaine aux applications nombreuses et prometteuses. Cependant, la recherche dans ce domaine est naissante et pas encore transférable dans le monde industriel. Ainsi, cette section ne présente pas de résultat concret, mais plutôt une description de stratégies existantes qui pourraient être appliquées sur un smartphone.

Notre étude s'appuie sur l'état de l'art des travaux de reconnaissance automatique de situations sociales par Vinciarelli et coll. (2009). La tâche a pour objectif l'identification automatique des comportements sociaux humains tels que les émotions, la personnalité, la relation de domination ou encore la persuasion. Pour cela, on cherche à reconnaître des indices sociaux liés à des traits du comportement :

— l'apparence physique,

— la gestuelle,

— les mouvements du visage et des yeux,

— la voix,

— la gestion de l'espace et de l'environnement.

Par exemple, les gestes des mains et la posture peuvent être employées pour la caractérisation de l'émotion d'une personne. Les auteurs précisent dans leur étude les principales techniques adoptées pour reconnaître les différents indices sociaux : l'analyse de la parole, l'analyse des images et la biométrie.

Le smartphone pourrait donc potentiellement être utilisé pour réaliser deux des trois techniques mentionnées, car l'appareil dispose d'un microphone et d'une caméra. Mais dans une utilisation quotidienne, le manque de contrôle de la caméra (suivant l'orientation ou la position, la capture peut être obstruée) et du microphone reste un problème pour l'instant peu abordé par la recherche.

Nous rapportons les travaux de Miluzzo et coll. (2008) qui proposent une application sur smartphone pour la reconnaissance de situations sociales. Les auteurs ne rapportent pas de résultat d'évaluation, mais nous décrivons le système qui est intéressant. Celui-ci embarque

un classifieur de voix humaine, qui fonctionne sur des fenêtres de signal courtes. L'identification de segments de voix déclenche un second traitement, opéré sur un serveur distant avec lequel l'application communique. Le traitement identifie une conversation au-delà d'un certain nombre de segments de voix reçus. Un troisième module est intégré, pour la reconnaissance de situations sociales prédéfinies telles que la présence d'amis, la présence d'une conversation ou l'environnement d'une fête. L'identification d'une situation combine le résultat du classifieur de conversation avec d'autres informations qui ne sont pas précisées dans l'article. Enfin, l'application tire profit du module de communication Bluetooth pour détecter les appareils à proximité et les identifier. L'hypothèse sous-jacente est que des personnes proches socialement (famille, amis, collègues) sont régulièrement rencontrées et peuvent laisser une trace *via* leur smartphone, notamment avec le Bluetooth activé.

2.2.6 Bilan des travaux de reconnaissance de contextes

Nous résumons dans la table 2.2 les tâches de reconnaissance d'éléments de contexte étudiées, regroupées suivant les types d'éléments de contexte identifiés et associées aux références bibliographiques des travaux étudiés.

Élément de contexte	**Description**	**Références**
Localisation utilisateur	Position sémantique (nom du lieu)	Wiki
	Points d'intérêt : détection et identification	Ravi *et al.*; Azizyan *et al.*
	Transitions : entrée/sortie bâtiment, modes de déplacement	Marmasse et Schmandt
Environnement	Événements sonores (parole, musique, autre),	Clarkson *et al.*
	Scènes sonores	Peltonen *et al.*; Ma *et al.*; Kern *et al.*
Activités physiques	Activités simples et postures	Bao et Intille; Incel *et al.*; Kwapisz *et al.*; Shoaib *et al.*
	Activités complexes	Bao et Intille; Dernbach *et al.*; Yan *et al.*
	Modes de transport (motorisés)	Sohn *et al.*; Reddy *et al.*
Contexte smartphone	Prise en compte de la position (main, oreille, poche, sac)	Bouten *et al.*; Hoseini-Tabatabaei *et al.*; Lester *et al.*
		Lane *et al.*; Lu *et al.*; Siirtola et Roning
	Reconnaissance de la position du smartphone	Blachon *et al.*; Diaconita *et al.*; Miluzzo *et al.*; Park *et al.*
Situations sociales	Détection de voix et de personnes	Vinciarelli *et al.* Miluzzo *et al.*

TABLE 2.2: Bilan des contextes étudiés en reconnaissance de contexte pour smartphone

De plus, pour chaque catégorie, nous dressons un bilan des éléments à retenir de cette section :

— la capacité de détecter les entrées et sorties du bâtiment par le GPS et l'identification de lieux par la "signature" des bornes Wi-Fi accessibles ;

— les travaux de reconnaissance d'environnements visent des situations proches de celles qui nous intéressent dans la thèse ; nous retenons les techniques qui sont éprouvées et les performances encourageantes ;

— les activités complexes étudiées (incluant les déplacements en transport motorisé) sont pertinentes pour les situations considérées dans la thèse et les résultats pour les tâches de reconnaissance de ces activités sont encourageants ; en outre, nous retenons les stratégies adoptées (combinaison de différentes sources, composition du système en plusieurs étapes) ;

— l'étude du contexte du smartphone a fourni des solutions pour le prendre en compte ou en tirer profit ;

— les travaux de reconnaissance de situations sociales ne sont pas encore fiables.

2.3 Les sources de données individuelles

Nous proposons une comparaison des sources de données mentionnées dans les travaux précédents. La première comparaison, plutôt qualitative, tente de déterminer la pertinence ou la redondance des sources suivant la diversité des tâches pour lesquelles elles sont employées et leur apport dans une tâche. La seconde comparaison est effectuée sur des critères objectifs liés à la faisabilité d'une tâche de reconnaissance sur smartphone.

2.3.1 Comparaison de l'usage des sources de données dans l'état de l'art

Nous résumons dans la table 2.3 les usages des sources dans les différentes tâches étudiées dans la section précédente. Nous avons retiré la tâche de localisation par récupération de labels sémantiques.

		Accéléromètre	*Gyroscope*	*Antennes-relais*	*Wi-Fi*	*GPS*	*Bluetooth*	*Microphone*
Env. Loc.	Points d'intérêt			✓	✓			
	Entrées sorties de bâtiments					✓		
	Environnements							✓
Ctxt. sp.	Activités physiques simples et postures	✓	✓					
	Activités complexes et transports	✓		✓		✓		
	Position du smartphone	✓						✓
	Orientation du smartphone	✓						
Social	Détection de personnes						✓	✓

TABLE 2.3: Matrice des distributions des sources pour la reconnaissance des éléments de contexte

L'analyse suivant les colonnes permet d'estimer à première vue les capteurs les plus populaires. Ainsi, l'accéléromètre est utilisé pour quatre tâches différentes, réparties dans deux catégories (la reconnaissance d'activités physiques et la reconnaissance du contexte du smartphone). Le microphone est impliqué dans trois tâches réparties dans autant de catégories. Ensuite, les antennes-relais et le GPS sont impliqués dans deux tâches réparties dans les deux mêmes catégories. Finalement, le gyroscope, le Wi-Fi et le Bluetooth sont employés dans une tâche unique d'après la revue des travaux. Cette observation met en avant l'accéléromètre, le microphone, les antennes-relais et le GPS, qui sont les plus populaires. À l'inverse, le gyroscope, le Wi-Fi et le Bluetooth sont moins employés.

L'analyse des lignes de la matrice est intéressante pour observer la diversité des sources pour certaines tâches. Commençons par la première ligne qui indique les sources d'antennes-relais et du Wi-Fi employées pour la reconnaissance de points d'intérêts. Toutes deux sont des sources de positionnement indirectes, dont la fonction première est la transmission de signaux. Pour chacune de ces sources, il est possible d'obtenir son identifiant ainsi que les coordonnées géographiques de son positionnement. Il est également possible d'estimer la puissance du signal reçu, afin d'estimer la distance à l'antenne par exemple. Les deux différences qui nous intéressent entre ces sources sont la précision (résultant de la portée) et la consommation de batterie. L'antenne-relais a une portée variant de quelques dizaines à plusieurs centaines de mètres suivant l'environnement de son implantation, contrairement à une borne Wi-Fi dont la portée ne dépasse pas quelques dizaines de mètres. Par conséquent, l'incertitude sur le positionnement à partir du Wi-Fi est moins importante. Cependant, qualitativement, l'usage du Wi-Fi pour détecter les bornes et communiquer nécessite beaucoup plus d'énergie que l'usage du réseau cellulaire qui permet de se connecter aux antennes-relais. Ces deux sources apparaissent donc complémentaires dans un premier temps.

La seconde ligne de la matrice indique que le GPS est la seule source mentionnée dans les travaux étudiés pour détecter les entrées et sorties de bâtiments. La même observation peut être faite du microphone pour la reconnaissance d'environnements. Nous passons directement à la tâche de reconnaissance d'activités physiques simples et de postures. Dans les travaux étudiés, l'accéléromètre est apparu comme la source incontournable. Il s'agit d'un capteur inertiel dont la valeur retournée est intuitive : l'accélération récupérée est comparable à une force projetée sur l'axe du capteur. Intuitivement, la mesure estime la quantité de mouvement appliquée. Le gyroscope fournit une mesure de la vitesse de rotation, qui est moins facile à identifier pour les activités et postures considérées. Cependant, l'étude de Shoaib et coll. (2014) a montré la complémentarité des deux sources.

Nous passons ensuite aux activités complexes et aux modes de transport. La complémentarité entre l'accéléromètre et les sources de positionnement d'antennes-relais et du GPS a été mise en évidence par les différentes études de reconnaissance de moyens de transport. L'accéléromètre peut distinguer l'immobilité, la marche ou le déplacement à vélo. Mais lorsqu'il s'agit d'un mode de transport, le GPS ou les antennes-relais sont plus efficaces. Habituellement, l'usage des modes de transport veut que l'on soit immobile, assis ou debout. Ainsi, les mouvements mesurés sont limités aux mouvements du véhicule, qui sont occasionnels et peu marqués, rendant difficile leur reconnaissance par les accélérations. Nous rappelons également que l'étude de Sohn et coll. (2006) avait montré de bons résultats pour la détection de la marche grâce aux antennes-relais.

Concernant la reconnaissance de la position du smartphone, l'accéléromètre et le microphone se sont révélés pertinents. La plupart des travaux relevés les ont employés séparément mais ont indiqué des résultats encourageants. Dans notre étude (2014a), nous avons combiné les descripteurs de ces deux sources et avons constaté une amélioration significative de la reconnaissance de la position. Ces deux sources apparaissent comme complémentaires.

Pour le calcul de l'orientation du smartphone, celui-ci n'a été effectué qu'à partir de l'accéléromètre dans les travaux rapportés, pour terminer avec la détection de situations sociales. L'usage du microphone a été employé pour la détection de voix et la reconnaissance de parole et de conversation. Le Bluetooth a également été évoqué dans une étude mais son usage est limité.

2.3.2 Comparaisons objectives des sources

Nous proposons dans la table 2.4 une comparaison des sources suivant des critères objectifs relativement à la tâche de reconnaissance d'éléments de contexte sur smartphone. Les sources étudiées sont celles déjà présentées dans les travaux de reconnaissance. Nous avons ajouté d'autres capteurs physiques présents sur un nombre grandissant d'appareils. Le capteur de luminosité donne une mesure de l'intensité. Le magnétomètre indique la mesure du champ magnétique suivant trois axes orthogonaux. Le baromètre mesure la pression atmosphérique. Le capteur de proximité est placé au-dessus de l'écran du smartphone et détecte la présence d'objets immédiatement proches. L'utilisation classique consiste à détecter la présence de l'oreille pendant un appel pour désactiver l'écran.

		Disponibilité	**Échantillonnage**	**Puis. instantanée (mW)**
Sources des travaux	Accéléromètre	Toujours	Régulier 50 Hz	5
	Gyroscope	Modérée	Régulier 50 Hz	**30**
	Antennes-relais	Toujours	Évenementiel	5-10
	Wi-Fi	Souvent	Évenementiel	**15-50**
	GPS	Souvent	Régulier 1 Hz	**380-430**
	Bluetooth	Souvent	Évenementiel	-
	Microphone	Toujours	Régulier 44100 Hz	-
Autres sources	Luminosité	Modérée	Régulier 50 Hz	3
	Magnétomètre	Souvent	Régulier 50 Hz	12
	Baromètre	Modérée	Régulier 50 Hz	1
	Proximité	Souvent	Régulier 50 Hz	7

TABLE 2.4: Comparaison des sources suivant des critères objectifs

Le critère le plus important est probablement la présence de la source sur l'appareil. Les smartphones modernes embarquent de plus en plus de sources et celles qui ont été rapportées dans les travaux précédents sont parmi les plus populaires sur les téléphones. Néanmoins, on note dans la table que certains capteurs physiques comme le gyroscope, le capteur de luminosité et le baromètre ont une présence modérée.

Nous présentons également l'échantillonnage des différentes sources considérées, et constatons que certaines sont régulières et d'autres événementielles. La plupart des sources régulières ont une fréquence d'échantillonnage faible, à l'exception du microphone qui peut atteindre des valeurs très élevées.

La consommation d'énergie est un autre critère important. Nous rapportons des résultats de mesure de consommation d'énergie effectués par Carroll et Heiser (2013). Brièvement, le

protocole expérimental, les mesures sont effectuées sur un smartphone Galaxy SIII, considéré par les auteurs comme un smartphone de gamme supérieure et à la pointe du marché au moment de sa sortie. Suivant des considérations techniques d'électronique, les auteurs ont identifié un composant à l'intérieur de l'appareil à partir duquel ils ont pu effectuer les mesures. Nous rapportons dans la table 2.4 les mesures de puissance.

Les valeurs de consommation de puissance de la table 2.4 sont très intéressantes par leur variabilité. Ainsi, le GPS est, de loin, la source qui consomme le plus. Les deux valeurs indiquées sont mesurées dans deux phases différentes. La première consiste en l'acquisition de la première position ; la seconde correspond à la phase de suivi de la position. Ensuite, le Wi-Fi affiche une consommation non négligeable. En particulier, ces valeurs sont données respectivement pour les états de veille du téléphone (à gauche) d'éveil, sans application qui fonctionne (à droite). Cependant, la transmission de données augmente sensiblement la consommation de la source Wi-Fi qui atteint alors les 400 mW dans un cas d'utilisation classique du navigateur d'Internet. Les deux valeurs de consommation relevées pour les antennes-relais représentent, à gauche, les antennes cellulaires de seconde génération (GSM) et, à droite, les antennes de troisième génération (3G). De plus, ces valeurs correspondent à un état de veille où la seule tâche consiste à maintenir la connexion à une antenne. Les cas de transmission des données *via* ces antennes augmentent la consommation d'énergie sensiblement plus que pour le Wi-Fi.

Les capteurs physiques sont les sources qui consomment le moins. Les valeurs sont faibles et les écrats-types donnés dans l'article de Carroll et Heiser (2013) le sont également. Cependant, on note la consommation du gyroscope, plusieurs fois supérieure à celle des autres capteurs physiques. En particulier, l'accéléromètre utilise six fois moins de puissance que le gyroscope, pour des tâches de reconnaissance similaires.

2.3.3 Bilan sur les sources de données

L'accéléromètre apparaît comme une source incontournable de par les performances des tâches dans lesquelles il est impliqué, son intégration à de nombreux appareils et sa très faible consommation d'énergie. Le microphone est également intéressant de par sa diversité d'utilisation et sa présence assurée sur un téléphone. Toutefois, la valeur de consommation manque pour affiner l'évaluation.

Le GPS se montre pertinent pour certaines tâches, mais sa très forte consommation incite à l'utiliser avec prudence. L'usage des antennes-relais a aussi montré des résultats pertinents et la consommation nécessaire à la maintenance de la connexion à une antenne est très faible, ce qui en fait une source intéressante. L'usage du Wi-Fi pour l'identification de lieu est intéressant et sa consommation relativement faible pour la maintenance d'une connexion.

Les autres sources ne paraissent pas pertinentes dans les tâches testées ou n'ont pas pu être testées.

2.4 Méthodes et algorithmes d'apprentissage automatique

Dans cette section, nous présentons plusieurs algorithmes et modèles employés dans le domaine de l'apprentissage et, pour certains, mentionnés dans les travaux de reconnaissance précédents. Des méthodes supervisées et non-supervisées sont présentées. Nous distinguons également la phase de sélection d'attributs, qui se situe en amont de l'entraînement d'un classifieur.

2.4.1 La sélection d'attributs

La sélection d'attributs vise à sélectionner un sous-ensemble d'attributs à partir de critères pré-définis. Witten et coll. (2005) justifient cette opération par l'impact négatif de la présence d'attributs non-pertinents sur les performances de classification. Par exemple, l'ajout d'un attribut binaire aléatoire à un corpus de données peut faire baisser la performance d'un arbre de décision de 5 à 10 % (2005). La sélection d'attributs est donc souhaitée pour réduire la "confusion" des classifieurs et ainsi augmenter la performance. Elle mène également à une réduction de la dimension des données, qui permet à l'expérimentateur de se concentrer sur les variables essentielles.

La sélection d'attributs est d'abord présentée dans son ensemble. Puis nous décrivons plus précisément l'usage des critères de gain d'information et de corrélation, qui offrent une description de la relation entre la classe et les attributs.

2.4.1.1 Quelques éléments théoriques sur la sélection d'attributs

Dash et Liu (1997) définissent la sélection d'attributs par la synthèse de plusieurs définitions issues de points de vue différents. Nous nous intéressons à la définition liée à l'approche de classification. Selon les auteurs, la sélection d'attributs consiste à déterminer un sous-ensemble d'attributs pour lequel la précision de classification ne diminue pas significativement.

À partir de cette définition, et de travaux étudiés, les auteurs ont proposé une procédure pour la sélection d'attributs, dont les étapes sont décrites dans la liste ci-dessous et illustrées dans la figure 2.4.

— une procédure de génération des sous-ensemble candidats ;

— une fonction d'évaluation du candidat ;

— un critère d'arrêt ;

— une procédure de validation du sous-ensemble candidat.

Trois catégories de méthodes de génération de sous-ensembles sont rapportées par Liu et coll. (2005). La première concerne les méthodes dites *complètes* qui se caractérisent par la garantie d'obtenir un sous-ensemble optimal. La solution la plus intuitive consiste à parcourir l'ensemble des combinaisons d'attributs, mais elle est aussi très coûteuse en ressources et en temps. C'est pourquoi des règles heuristiques sont employées pour réduire l'es-

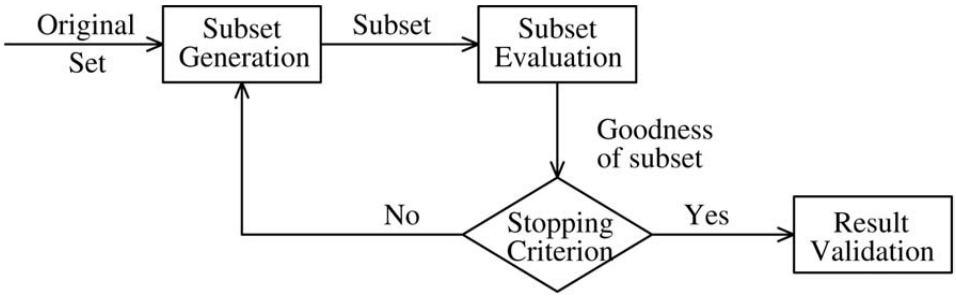


FIGURE 2.4: Illustration du processus de sélection d'attributs, extrait de l'article de Dash et Liu (1997)

pace des combinaisons à explorer tout en garantissant l'optimalité de la solution trouvée. La seconde catégorie de méthodes regroupe les méthodes dites *heuristiques*. Elles consistent à appliquer des règles heuristiques sur l'ensemble des combinaisons en vue de réduire sa taille. Cependant, contrairement aux méthodes de la première catégorie, elles ne garantissent pas l'optimalité globale de la solution. Ainsi, les règles appliquées sont plus contraignantes, ce qui permet de fortement diminuer le nombre de solutions et donc le temps et les ressources de traitement. La méthode dite séquentielle vers l'avant (*Sequential forward selection* en anglais) est un exemple de génération heuristique. Le principe repose sur une progression linéaire dans l'espace des combinaisons. La méthode est itérative : à chaque tour, un attribut est ajouté à l'ensemble existant suivant un critère qui estime son apport à l'ensemble des attributs. La méthode démarre avec un ensemble vide et s'arrête lorsque tous les attributs ont été ajoutés. La troisième catégorie de génération de sous-ensemble représente les méthodes aléatoires. Constatant les limites des méthodes des deux catégories précédentes dus au compromis entre optimalité et rapidité de résultat, l'idée est venue de choisir aléatoirement les combinaisons d'attributs. La méthode est itérative : une combinaison est sélectionnée aléatoirement à chaque itération.

L'évaluation des combinaisons générées offre également de nombreuses fonctions suivant qu'elles considèrent un sous-ensemble ou les attributs seuls, suivant aussi que les critères dépendent de la tâche finale ou non. Par exemple, dans la suite, nous décrivons l'évaluation par ratio de gain d'information, qui classe les attributs individuellement. Nous présentons aussi une méthode basée sur la corrélation qui traite des sous-ensembles d'attributs. La sélection des sous-ensembles est faite par la procédure dite *Sequential forwarding selection*, qui augmente la taille du sous-ensemble de manière itérative en choisissant l'attribut qui permet la plus forte augmentation de score suivant le critère considéré. Ces deux méthodes reposent sur des critères dits objectifs, mais il existe aussi des méthode d'évaluation qui dépendent de la tâche finale. C'est par exemple le cas lorsque l'on effectue une sélection d'attributs avec un classifieur qui servira ultérieurement à la tâche de classification. Ce genre de méthodes permet d'obtenir une sélection plus adaptée à la tâche définie, mais convient moins à l'exploration générale.

2.4.1.2 L'évaluation par le ratio du gain d'information et la corrélation

Nous décrivons précisément les fonctions d'évaluation du gain d'information et de la corrélation car ces deux méthodes offrent une description de la relation entre la classe et les attributs, ce qui est particulièrement intéressant dans notre cas pour relier les attributs aux scènes.

Le gain d'information

Dans la théorie de l'information, le gain d'information (*Information Gain* en anglais) mesure la différence d'information sur une variable résultant de l'observation d'une autre variable. L'information est exprimée par l'entropie de la variable et le gain d'information par la différence entre l'entropie *a priori* d'une variable et l'entropie conditionnelle de cette même variable connaissant une seconde variable (Hall et Holmes (2003)). Si l'on considère C la classe et A un attribut, alors le gain d'information par l'observation de l'attribut A peut s'exprimer suivant la formule de l'équation 2.1. Nous rappelons les formules de calcul de l'entropie et de l'entropie conditionnelle dans les équations 2.2 et 2.3, nécessaires à l'expression du gain d'information.

$$IG = H(C) - H(C|A) \tag{2.1}$$

$$H(C) = -\sum_{c \in C} p(c) \log_2 p(c) \tag{2.2}$$

$$H(C|A) = -\sum_{a \in A} p(a) \sum_{c \in C} p(c|a) \log_2 p(c|a) \tag{2.3}$$

Cependant, le calcul du gain d'information est influencé par le domaine des valeurs des attributs. Pour éviter ce biais, notamment dans la comparaison d'attributs catégoriques et numériques, il est possible de normaliser le gain avec l'entropie de l'attribut. On parle alors de ratio de gain d'information (*gain ratio, GR*) :

$$GR = \frac{IG}{H(A)} = \frac{H(C) - H(C|A)}{H(A)} \tag{2.4}$$

L'évaluation par corrélation

La méthode de sélection par corrélation (*Correlation-based Feature Selection*) a été proposée par Hall (1999). Elle évalue les sous-ensembles d'attributs (contrairement au calcul du gain d'information qui s'effectue sur les attributs individuels). La méthode calcule un score de mérite, qui exprime la pertinence d'un sous-ensemble pour la prédiction de la classe. Le mérite est le ratio de deux éléments. Le numérateur exprime la corrélation des attributs du sous-ensemble avec la classe. Le dénominateur exprime la redondance des attributs dans le sous-ensemble, mesurée par l'inter-corrélation entre les attributs. Ainsi, le mérite sera élevé pour un ensemble d'attributs corrélés à la classe et peu redondants. À l'inverse, un ensemble hautement corrélé à la classe mais avec une forte redondance sera sanctionné d'un mérite faible. De même, un ensemble faiblement corrélé à la classe mais peu redondant aura un mérite faible.

L'équation 2.5 décrit le calcul du mérite. La variable k représente le nombre d'attributs du

sous-ensemble. Le terme $\overline{r_{cf}}$ exprime la moyenne des corrélations calculées entre la classe et chaque attribut du sous-ensemble. La corrélation est calculée en appliquant la formule du coefficient de Pearson. Les équations 2.6 et 2.7 précisent le calcul du coefficient, respectivement pour des variables continues X et Y et pour la combinaison de variables discrète C et continue Y. Le terme $\overline{r_{ff}}$ représente l'inter-corrélation moyenne entre les attributs.

$$\text{Mérite}_s = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \tag{2.5}$$

$$r_{XY} = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y} \tag{2.6}$$

$$r_{CY} = \sum_{i=1}^{k} p(C = c_i) r_{C_i Y} \tag{2.7}$$

2.4.2 Modèles et algorithmes d'apprentissage automatique

La section décrit cinq algorithmes ou modèles d'apprentissage automatique mentionnés dans les travaux de reconnaissance d'éléments de contexte.

2.4.2.1 Réseau bayésien et réseau bayésien naïf

Pour définir un réseau bayésien, considérons un ensemble fini de variables aléatoire $\Omega_X = \{X_1, \ldots, X_n\}$, chacune pouvant prendre une valeur x_i dans un ensemble de valeurs associé Ω_{X_i}. Un réseau bayésien est composé d'un graphe orienté acyclique (*directed acyclic graph* en anglais) dont les nœuds correspondent aux variables aléatoires de Ω_X et dont les arcs entre les nœuds définissent les dépendances entre les variables associées.

Le graphe possède une propriété particulière qui indique l'indépendance conditionnelle d'une variable X_i à toute autre variable, connaissant ses parents dans le graphe. Cette hypothèse permet de simplifier l'expression de la probabilité conjointe de toutes les variables du graphes à l'expression suivante :

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{N} P(X_i | pa(X_i)) \tag{2.8}$$

où $pa(X_i)$ est l'ensemble des parents de X_i dans le graphe G.

Nous illustrons la définition avec l'exemple d'un réseau bayésien dans la figure 2.5, repris de l'article de Pearl (2011).

Le graphe exprime la réalisation de plusieurs événements, représentés par des variables aléatoires, et structurés par des connaissances ou hypothèses représentées par les arcs. D'après la propriété d'indépendance conditionnelle d'un nœud aux autres nœuds que ses parents, la probabilité d'une *glissade* (variable X_5) est conditionnellement dépendante de la probabilité que le sol soit *mouillé* (variable X_4). Cette dernière variable est conditionnellement dépendante des états de l'*arrosage automatique* (variable X_3) et de la *pluie* (variable X_2). Globalement, pour calculer la probabilité jointe de toutes les variables, on reprend la

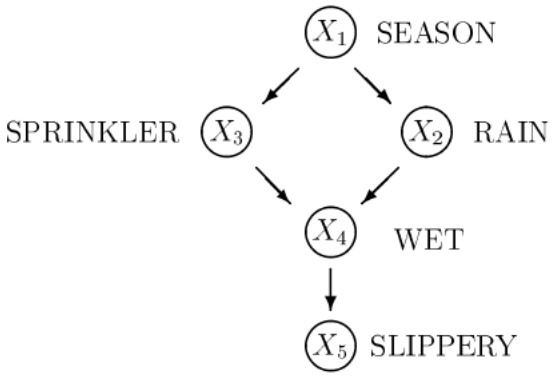


FIGURE 2.5: Exemple de réseau bayésien extrait de l'article de Pearl (2011)

formule de l'équation 2.8 :

$$P(X_1, X_2, X_3, X_4, X_5) = P(X_1)P(X_2|X_1)P(X_3|X_1)P(X_4|X_2, X_3)P(X_5|X_4) \tag{2.9}$$

Le réseau bayésien naïf est un cas particulier des réseaux bayésiens car une hypothèse supplémentaire est formulée. Pour cela, considérons un problème de classification avec d'une part, une variable aléatoire C qui prend ses valeurs dans un ensemble fini $(c_1, \ldots, c_p)$ des classes possibles ; et d'autre part un ensemble de variables aléatoires $(X_1, \ldots, X_n)$ qui représentent les descripteurs issus des observations. L'hypothèse exprime la dépendance directe des variables d'observations à la variable de la classe. De plus, l'application d'un réseau bayésien naïf à un problème de classification vise à déterminer la valeur de la classe qui est la plus probable en connaissant les variables d'observation. Cette probabilité est exprimée par la quantité $P(C = |X_1, \ldots, X_n)$. La règle de Bayes est employée pour la calculer :

$$P(C|X_1, \ldots, X_n) = \frac{P(X_1, \ldots, X_n|C)P(C)}{P(X_1, \ldots, X_n)} \tag{2.10}$$

L'équation peut être simplifiée grâce à l'hypothèse d'indépendance conditionnelle d'une variable aux autres variables que ses parents qui permet d'écrire l'égalité suivante :

$$P(X_1, \ldots, X_n|C) = \prod_{i=1}^{n} P(X_i|C) \tag{2.11}$$

et, par suite :

$$P(C|X_1, \ldots, X_n) = \frac{\prod_{i=1}^{N} P(X_i|C)P(C)}{P(X_1, \ldots, X_n)} \tag{2.12}$$

On peut aussi remarquer que la probabilité *a priori* $P(X_1, \ldots, X_n)$ est invariante pour l'évaluation de la quantité $P(C|X_1, \ldots, X_n)$ suivant les différentes classes. Ainsi, la détermination de la classe la plus probable peut être limitée à la maximisation du dénominateur de

l'équation 2.12 de la manière suivante :

$$\hat{c_k} = \underset{c=c_1,\ldots,c_p}{\operatorname{argmax}}\big(P(C=c|X_1,\ldots,X_n)\big) = \underset{c=c_1,\ldots,c_p}{\operatorname{argmax}}\big(\prod_{i=1}^{N} P(X_i|C=c)P(C=c)\big) \tag{2.13}$$

Les probabilités $P(X_i|C)$ sont établies pendant la phase d'entraînement du classifieur. En pratique, l'hypothèse d'indépendance conditionnelle ne peut être toujours vérifiée, mais le réseau bayésien naïf représente souvent une référence difficile à battre. En outre, les probabilités *a priori* des classes $P(C)$ peuvent être estimées empiriquement, sur le corpus de données.

2.4.2.2 Les arbres de décision

Nous présentons d'abord le fonctionnement de l'arbre de décision de la méthode C4.5. Puis, nous abordons la technique dite du *bagging* appliquée aux arbres et qui constitue l'algorithme dit de *Random Forest*.

L'arbre de décision

Un arbre de décision est un graphe direct acyclique dont les nœuds constituent des tests à effectuer sur les attributs, et les arcs sont des intervalles de valeurs pour les tests. Plusieurs algorithmes existent pour la construction du graphe. Nous présentons le C4.5 (Quinlan, 1993), qui est une version très populaire. La construction du graphe est effectuée progressivement par l'évaluation des attributs (ou descripteurs) sur le corpus de données et la sélection du test qui apporte le plus d'information sur l'organisation du corpus. Nous présentons le pseudo-code de la création de l'arbre ci-dessous.

Pour chaque descripteur d, calculer le gain d'information par le test effectué sur d ;
Sélectionner d_{best} qui offre le gain d'information le plus élevé ;
Créer un nœud avec le test sur d_{best} ;
Répartir les données suivant les valeurs prises par le test sur d_{best} ;
Recommencer pour chaque sous-ensemble de données ;

L'algorithme C4.5 emploie le ratio de gain d'information comme test. Comme expliqué précédemment, le gain d'information mesure la quantité d'information apportée sur une classe par la connaissance d'un attribut. Le ratio du gain normalise la mesure par l'information présente dans l'attribut. L'application à l'arbre de décision consiste à comparer l'apport d'information de chaque attribut et à sélectionner le maximum. Intuitivement, un fort gain d'information indique une importante entropie conditionnelle de la classe relativement à l'attribut, ce qui justifie la division du corpus suivant les intervalles de valeurs de cet attribut.

L'arbre de décision offre plusieurs avantages. La représentation des données est lisible et facile à interpréter, ce qui permet d'utiliser un arbre de décision non seulement pour la classification, mais aussi pour l'exploration de données, par exemple pour déterminer les

attributs les plus pertinents dans un ensemble de données. Ensuite, l'arbre de décision peut traiter des données numériques et catégorielles. Également, le processus pour l'inférence, qui consiste en une succession de test, est très rapide, ce qui le rend pertinent pour un usage où le flux de données est continu ou avec de grands corpus de données. Cependant, l'arbre compte plusieurs défauts. Le premier est le manque de généralisation, inhérent à l'algorithme qui tend à considérer des ensembles de données de plus en plus réduits. Des mécanismes tels que l'élagage sont proposés pour arrêter l'exploration de l'algorithme lorsque le sous-ensemble considéré est trop faible. Ensuite, l'algorithme d'apprentissage évolue par la recherche successive de *maxima* locaux de gains d'informations pour déterminer les tests. Cette approche est préférée à la construction d'un arbre de décision optimal, beaucoup plus coûteuse en temps et ressources, au détriment d'un graphe *a priori* sous-optimal.

La forêt d'arbres décisionnels (*Random Forest*)

Breiman (2001) définit la forêt d'arbres décisionnels comme "*un classifieur composé d'un ensemble d'arbres de décision* $\{h(x,\Theta_k), k = 1,\ldots,N\}$ *où les* $\{\Theta_k\}$ *sont des ensembles du corpus d'entraînement indépendants et uniformément distribués ; chaque arbre propose une classe pour évaluer l'échantillon* x *et la classe la plus probable est déterminée par un vote sur les propositions de tous les arbres*".

L'entraînement de la forêt d'arbres décisionnels repose sur deux techniques. La première, appelée *bagging* en anglais (contraction de *bootstrap aggregating*), tire profit d'un corpus de données pour entraîner un nombre N de classifieurs sur des portions du corpus et à combiner les prédictions de tous les classifieurs pour un échantillon inconnu par vote majoritaire si la classe est de type catégoriel ou par moyenne si elle est numérique (Breiman, 1996). De plus, les portions du corpus d'entraînement sont constituées par échantillonnage aléatoire avec remplacement, afin de les considérer comme indépendants. Cette méthode permet d'obtenir des arbres décorrélés, ce qui constitue l'une des forces de l'algorithme.

La seconde technique consiste à considérer pour l'entraînement de chaque arbre un sous-ensemble des descripteurs choisi aléatoirement. Cette technique est justifiée pour limiter la corrélation entre les arbres. En effet, si l'on imagine que quelques descripteurs ont un pouvoir discriminant très fort, alors il est probable qu'ils seront sélectionnés dans l'entraînement de nombreux arbres, menant potentiellement à une augmentation de la corrélation entre les arbres.

L'un des avantages de cet algorithme est la faible tendance au sur-apprentissage (comparativement à un arbre de décision seul) (Breiman, 2001). L'usage de la forêt aléatoire s'est montré très efficace sur de grands jeux de données et a conduit à de très bons scores sur de nombreux corpus (comme dans la reconnaissance d'activités physiques (Cvetkovic, 2013)).

2.4.2.3 Les mélanges de gaussiennes (GMM)

Le GMM (ou mélanges de gaussiennes) est une extension du modèle gaussien qui tente de représenter une distribution de probabilités par la somme pondérée de plusieurs gaussiennes. Nous rappelons dans un premier temps la définition du modèle gaussien puis nous

introduisons les mélanges de gaussiennes.

Le modèle gaussien

Le modèle gaussien consiste à représenter la distribution de probabilité d'une variable aléatoire X par une distribution gaussienne $\mathcal{N}(\mu,\sigma)$ de paramètres μ sa moyenne et σ sa variance. L'expression de la probabilité de réalisation d'une variable aléatoire X qui suit cette loi est la suivante :

$$p(X=x)=\frac{1}{\sqrt{2\pi\sigma}}e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma}} \tag{2.14}$$

Si l'on considère maintenant le vecteur X comme la variable aléatoire qui suit une distribution gaussienne $\mathcal{N}(\overline{\mu},\Sigma)$ de paramètres $\overline{\mu}$ le vecteur moyen et Σ la matrice de covariance, alors l'équation précédente devient :

$$p(X=\overline{x})=\frac{1}{(2\pi)^{d/2}\sqrt{\det\Sigma}}e^{-\frac{1}{2}(\overline{x}-\overline{\mu})^T\Sigma^{-1}(\overline{x}-\overline{\mu})} \tag{2.15}$$

Dans le cas de la classification, une classe est associée à un modèle. Le modèle le plus vraisemblable est sélectionné par la probabilité de réalisation la plus élevée :

$$\hat{c}=\underset{c\in C}{\operatorname{argmax}}\ p_c(X=\overline{x})=\underset{c\in C}{\operatorname{argmax}}\left(\frac{1}{(2\pi)^{d/2}\sqrt{\det\Sigma_c}}e^{-\frac{1}{2}(\overline{x}-\overline{\mu_c})^T\Sigma_c^{-1}(\overline{x}-\overline{\mu_c})}\right) \tag{2.16}$$

Les mélanges de gaussiennes

Bien souvent, les distributions de probabilité des variables aléatoires ne suivent pas qu'une simple gaussienne. Une solution consiste à estimer la distribution souhaitée par une somme pondérée de gaussiennes. Le mélange est alors défini par les paramètres suivants :

— P le nombre de gaussiennes ;

— π_p le poids associé à chaque distribution gaussienne ;

— $\overline{\mu}_p$ le vecteur moyen de la gaussienne d'indice p ;

— Σ_p la matrice de covariance associée à la gaussienne d'indice p.

La probabilité de réalisation d'une variable aléatoire X s'exprime en reprenant l'équation 2.15 par la somme pondérée des distributions individuelles :

$$p(X=\overline{x})=\sum_{p=1}^{P}\pi_p\frac{1}{(2\pi)^{d/2}\sqrt{\det\Sigma_p}}e^{-\frac{1}{2}(\overline{x}-\overline{\mu}_p)\Sigma_p^{-1}(\overline{x}-\overline{\mu_p})} \tag{2.17}$$

L'entraînement consiste en l'estimation des paramètres $\{\pi_p,\overline{\mu}_p,\Sigma_p|1\leq p\leq P\}$. Il est réalisé par l'algorithme *Expectation Maximization* (*EM*) décrit par Bilmes (1998) comme :

> *une méthode générale pour l'estimation par maximisation de la vraisemblance des paramètres d'une distribution sous-jacente à partir d'un ensemble de données, qui peut être incomplet ou manquer certaines données.*

La méthode considère l'ensemble X des données observées et supposées générées par une distribution. Celui-ci est considéré incomplet et suppose l'existence d'un ensemble de

données complet représenté par $Z = (X, Y)$ où Y est une variable supposée inconnue, aléatoire et produite par une distribution sous-jacente. L'ensemble Z peut être relié à l'ensemble Θ des paramètres à estimer par une probabilité jointe :

$$p(z|\Theta) = p(x, y|\Theta) = p(y|x, \Theta)\, p(x|\Theta) \tag{2.18}$$

L'algorithme est itératif et procède en deux étapes au cours d'une itération. La première étape consiste à calculer l'espérance de la log-vraisemblance du modèle pour l'ensemble de données complet $\log L(\Theta|X, Y)$ relativement à la variable aléatoire Y connaissant les données observées X et les valeurs courantes estimées des paramètres du modèle Θ :

$$Q\left(\Theta, \Theta^{(i-1)}\right) = E\left[\log p(X, Y|\Theta)|X, \Theta^{(i-1)}\right] \tag{2.19}$$

Dans cette expression, X et $\Theta^{(i-1)}$ sont constantes et Y suit la distribution décrite par $f(y|X, \Theta^{(i-1)})$. De plus, le développement de l'espérance permet de récrire l'équation de la manière suivante :

$$E\left[\log p(X, Y|\Theta)|X, \Theta^{(i-1)}\right] = \int_{y \in \Upsilon} \log p(X, y|\Theta) f(y|X, \Theta^{(i-1)})\, dy \tag{2.20}$$

Une fois la quantité précédente évaluée, la seconde étape vise à maximiser la quantité précédente suivant les paramètres du modèle Θ :

$$\Theta^{(i)} = \underset{\Theta}{\operatorname{argmax}}\, Q\left(\Theta, \Theta^{(i-1)}\right) \tag{2.21}$$

La vraisemblance est assurée d'augmenter à chaque itération et de converger vers un maximum local. Une version modifiée de l'étape de maximisation consiste à rechercher un ensemble de paramètres $\Theta^{(i)}$ qui permet l'augmentation de la quantité Q de sorte que l'on ait $Q\left(\Theta^{(i)}, \Theta^{(i-1)}\right) > Q\left(\Theta, \Theta^{(i-1)}\right)$. Cette forme particulière est appelée l'algorithme EM généralisé (GEM) et elle garantit aussi la convergence de la vraisemblance vers un maximum local.

2.4.2.4 Les réseaux de neurones artificiels et les réseaux de neurones profonds

Le DNN (*Deep Neural Network* en anglais) est le réseau de neurones qui nous intéresse. Cependant, afin de simplifier sa description, nous présentons d'abord le perceptron, un réseau à un neurone, puis nous décrivons les réseaux à plusieurs couches.

Le perceptron

Le perceptron est le réseau de neurones élémentaire. Il est composé d'un seul neurone, connecté d'un côté directement à des entrées lui fournissant des données et de l'autre à une sortie à laquelle il fournit le résultat du traitement.

Le neurone du perceptron, tout comme les neurones de réseaux plus complexes, réalise la transformation des signaux d'entrée en deux temps. La figure 2.6 représente son fonc-

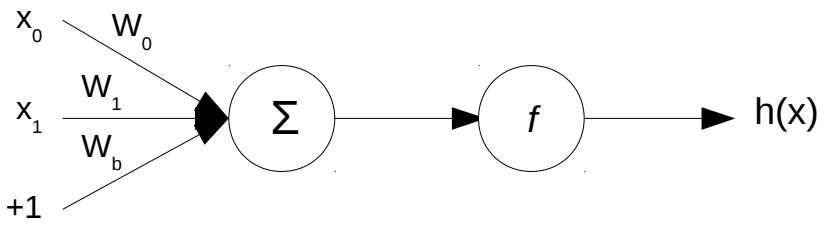


FIGURE 2.6: Illustration d'un perceptron

tionnement. Dans cet exemple, deux entrées sont soumises au perceptron, pondérées par les valeurs W_0, W_1, puis sommées. On remarque également une entrée de valeur 1, il s'agit d'un *biais*. Comme les autres entrées, il est associé à un poids W_b et intégré à la somme. Le résultat de la somme est soumis à la fonction d'activation f qui génère la sortie $h(x)$ du neurone. La fonction d'activation (ou seuil d'activation) représente la "stimulation" du neurone à l'information reçue en entrée. La fonction sigmoïde est un exemple courant de fonction d'activation.

L'entraînement du perceptron vise à déterminer les valeurs des poids W_i qui minimisent l'erreur de classification. Pour cela, un corpus d'entraînement est constitué de l'ensemble des vecteurs d'entrées associés aux sorties produites par le perceptron et aux valeurs réelles attendues (les "classes"). Si l'on considère les notations suivantes :

— $x = (a_0, \ldots, a_N)$ un vecteur d'entrée (dans l'exemple de la figure, N vaut 1),

— W_j le poids du $j^{ième}$ arc entrant du neurone, incluant le poids du biais W_b,

— $\{x_1 \rightarrow y_1, \ldots, x_M \rightarrow y_M\}$ un corpus de M exemples,

— $h(x)$ les sorties calculées pour le vecteur x,

alors on peut exprimer la sortie du perceptron en fonction de l'entrée par la formule suivante :

$$h(x) = f\left(\sum_i W_i a_i\right) \tag{2.22}$$

et par suite, on peut exprimer l'erreur quadratique E^2 qui servira à l'estimation des poids :

$$E^2 = (y - h(x))^2 \tag{2.23}$$

L'estimation des poids est itérative et utilise la méthode du gradient descendant. À chaque itération, les dérivées partielles de l'erreur sont calculées relativement à chaque poids et représentent le pas qui servira d'avancement au gradient. La dérivée partielle de l'erreur s'exprime suivant la formule :

$$\frac{\partial E^2}{\partial W_i} = E.\frac{\partial E}{\partial W_i} = E.\frac{\partial\left(y - f\left(\sum_k W_k a_k\right)\right)}{\partial W_i} \tag{2.24}$$

$$= -E.f'(\sum_k W_k a_k).a_i \tag{2.25}$$

Par suite, cette expression est utilisée pour mettre à jour les valeurs des poids suivant la méthode du gradient descendant, où α est le taux d'apprentissage :

$$W_i \leftarrow W_i + \alpha.E.f'(\sum_k W_k a_k).a_i \qquad \text{avec } \alpha \in]0,1[\tag{2.26}$$

Le perceptron est très adapté pour les problèmes linéaires mais ne peut qu'approximer les problèmes de classification non-linéaires.

Les réseaux de neurones multi-couches

Afin de permettre la représentation d'un plus grand nombre de fonctions, le perceptron est enrichi d'autres neurones répartis en couches inter-connectées : c'est le réseau de neurones multi-couches. Un tel réseau contient un minimum de 3 couches, à l'image de celui représenté sur la figure 2.7. Une couche d'entrée représente les valeurs du vecteur d'entrée. Une couche de sortie produit les sorties du réseau. Une ou plusieurs couches intermédiaires, aussi appelées couches *cachées,* réalisent des transformations. On remarque les biais introduits pour le perceptron. Les biais sont ajoutés à la couche d'entrée et à chaque couche intermédiaire.

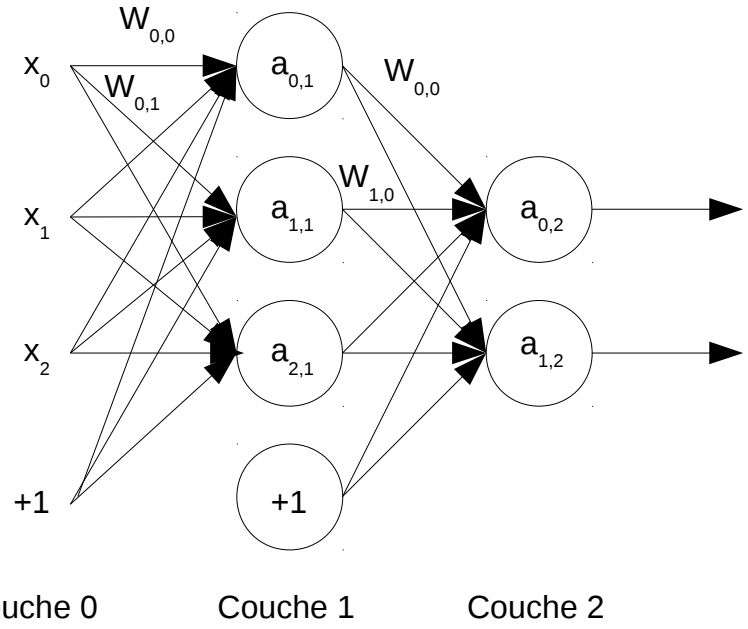


FIGURE 2.7: Illustration d'un réseau de neurones multi-couches

Comme pour le perceptron, l'entraînement du réseau consiste à minimiser les poids des arcs du réseau. La méthode du *gradient descendant* est appliquée à partir du calcul de l'erreur quadratique. La mise à jour des poids de la dernière couche est directement possible à partir de la formule 2.26. Les poids des couches intermédiaires ne peuvent être dérivés directement du calcul de l'erreur sur la couche de sortie. Cependant, il est possible d'estimer la contribution d'un neurone à l'erreur globale par la contribution de ses arcs sortants à la couche supérieure et ainsi, d'estimer la part de l'erreur pour ce neurone. L'estimation du

poids d'une couche intermédiaire se fait avec la formule suivante :

$$W_{i,j} \leftarrow W_{i,j} + \alpha . f'\left(\sum_k W_{k,j} a_k\right) . \sum_k W_{j,k} \Delta k \tag{2.27}$$

où Δ_k représente l'erreur de chaque neurone de la couche supérieure, $W_{j,k}$ les poids des arcs issus du neurone courant vers les neurones de la couches supérieure et $f'\left(\sum_k W_{k,j} a_k\right)$ la dérivée de la fonction d'activation du neurone courant relativement à la somme pondérée de ses entrées. Intuitivement, cette formule ressemble à la formule valable pour la couche de sortie. La différence réside dans l'expression de l'erreur, qui est pondérée par l'importance de la contribution du neurone, elle-même représentée par la pondération des arcs du neurone vers les nœuds de sortie.

Les réseaux de neurones profonds

L'expression de "réseaux de neurones profonds" fait référence au nombre de couches cachées du réseau (supérieur à 1) et à la méthode d'apprentissage profond (*deep learning*) employée pour entraîner ce réseau. En effet, la méthode d'apprentissage par gradient descendant présentée précédemment combinée à une initialisation aléatoire des poids est limitée à un optimum local lorsqu'elle est appliquée à l'entraînement de réseaux de neurones profonds (Bengio, 2009).

Pour résoudre ce problème, Hinton et coll. (2012) ont proposé une méthode d'apprentissage qui consiste d'abord à entraîner individuellement et successivement les couches du réseau (phase de pré-entraînement ou d'initialisation), puis à appliquer la méthode précédente de gradient descendant pour optimiser les paramètres initialisés. Pour l'entraînement d'une couche, les auteurs suggèrent l'usage de machines de Boltzmann réduites (*Restricted Boltzmann Machines* en anglais, abrégées par RBM). Il s'agit d'un réseau de neurone artificiel composé d'une couche visible en entrée et d'une couche cachée et qui peut apprendre une distribution probabiliste. La méthode proposée par Hinton et coll. applique d'abord une RBM sur la couche d'entrée des données brutes pour la modéliser. La couche cachée de la RBM (qui modélise donc les données) est ensuite employée comme couche d'entrée dans l'entraînement de la couche suivante du réseau de neurones profonds global. On procède ainsi pour l'entraînement de toutes les couches cachées du réseau de neurones profonds. À l'issue, on reconstruit le réseau par "l'empilement" des couches et l'ajout de la couche de sortie qui permet d'exprimer la probabilité d'obtenir chacune des classes possibles (pour plus de détails, se référer à l'article de Hinton et coll. (2012)). L'empilement des RBMs entraînés successivement est défini comme un DBN (*Deep Belief Net* en anglais).

2.4.3 Méthodes d'analyse non-supervisée

Nous décrivons deux méthodes d'analyse non-supervisée. Elles n'ont pas été mentionnées dans l'état de l'art, mais sont employées dans les expérimentations. La première méthode traite la séquence d'instances pour les représenter progressivement en segments ho-

mogènes. La seconde méthode est un algorithme de regroupement hiérarchique probabiliste basé sur le principe de l'algorithme EM introduit précédemment.

2.4.3.1 Méthode de segmentation de série temporelle

De manière générale, la segmentation vise à diviser une séquence ordonnée d'éléments en segments finis afin de mettre en évidence des motifs particuliers. Nous présentons une méthode de segmentation de série temporelle initialement développée pour la segmentation en locuteurs d'un enregistrement sonore (Le et coll. (2007)). Nous considérons que les conditions de fonctionnement de la méthode pour cette tâche (en particulier, durée et identification des tours de paroles inconnues) sont semblables à celles de la reconnaissance de scène (durée et identification d'une scène inconnues). C'est pour cette raison que nous avons choisi cette méthode.

L'approche est non-supervisée et se compose de trois étapes principales : la détection de changements qui représentent les frontières des segments ; la combinaison des segments adjacents ; et le regroupement hiérarchique des segments non contigus. D'autres étapes peuvent être ajoutées mais nous ne les utilisons pas dans le travail de ce manuscrit.

Notre description représente la méthode telle qu'elle est appliquée dans les expérimentations. En particulier, les modèles de représentation sont gaussiens et les critères reposent sur l'estimation de la vraisemblance du modèle.

Détection des changements de concept

La première étape de segmentation traite le flux de vecteurs progressivement, en appliquant une fenêtre glissante contenant un nombre M de vecteurs. Le but est de déterminer si la fenêtre contient un changement de concept et nécessite la pose d'une frontière. Deux hypothèses sont testées :

— l'hypothèse H_0 suppose que l'ensemble des vecteurs de la fenêtre est issu du même modèle ;

— l'hypothèse H_1 suppose que le groupe de vecteurs de la moitié gauche de la fenêtre et celui de la moitié droite sont issus de deux modèles différents.

La vérification de ces hypothèses requiert l'estimation de trois modèles : un modèle λ_0 qui représente l'hypothèse H_0 et deux modèles λ_{11} et λ_{12} qui représentent l'hypothèse H_1. Les paramètres de moyenne μ et matrice de covariance Σ des modèles gaussiens multivariés sont estimés sur les vecteurs des différentes fenêtres introduites. La vérification des hypothèses consiste à évaluer celle qui est la plus vraisemblable. La vraisemblance qu'un modèle ait généré une séquence d'observations est représentée par la probabilité *a posteriori* $p(x|\lambda)$ de la séquence d'observations x connaissant le modèle λ. Pour l'hypothèse H_0, la vraisemblance du modèle λ_0 pour la séquence d'observations x s'exprime par l'écriture $L\left(\mu_0, \Sigma_0|\lambda_0\right)$. La vraisemblance globale de l'hypothèse H_1 est calculée par le produit des vraisemblances des modèles λ_{11} et λ_{12}. L'évaluation des vraisemblances est réalisée par la différence des

log-vraisemblances, exprimée par la formule 2.29 :

$$R(H_0, H_1) = -\log\left(\frac{L_0}{L_1}\right) \tag{2.28}$$

$$= -\log L\left(\mu_0, \Sigma_0 | x\right) + \log L\left(\mu_{11}, \Sigma_{11} | x_1\right) + \log L\left(\mu_{12}, \Sigma_{12} | x_2\right) \tag{2.29}$$

La décision de segmenter est confirmée par le signe positif de l'expression qui indique la vraisemblance plus forte de l'hypothèse H_1. Dans la pratique, il arrive que le signe du rapport soit positif sur des fenêtres successives, qui sont considérées comme des zones d'incertitude. Ainsi, la règle n'est pas appliquée strictement. À la place, on recherche localement le point pour lequel le rapport est positif et le plus élevé pour appliquer la frontière.

Combinaison des segments adjacents

Par expérience, le résultat de la segmentation est souvent fortement morcelé, composé de nombreux segments de durée très courte et résultant en une segmentation loin de ce qui était attendu. Plusieurs raisons peuvent expliquer cela (l'usage de modèles simples comme la distribution gaussienne, la taille limitée des fenêtres pour l'estimation ou encore la position arbitraire des fenêtres sur le signal). Afin de pallier ce problème, l'étape de segmentation est appliquée une seconde fois. La différence avec l'étape précédente réside dans les fenêtres employées qui correspondent aux segments détectés. Ainsi, la fenêtre pour la vérification des deux hypothèses H_0 et H_1 est composée de deux segments contigus pris successivement sur la séquence des segments. Ces derniers ne sont pas forcément de taille identique pour l'estimation des modèles. Afin d'intégrer ce point, le score d'évaluation est adapté en complétant la log-vraisemblance par l'ajout d'une pondération en fonction de la taille de la fenêtre considérée, c'est le critère BIC (*Bayesian Information Criterion*, ou critère d'information bayésienne en français). Le critère d'évaluation des hypothèses évolue aussi : il calcule la différence des deux scores BIC. Les équations suivantes expriment le critère BIC pour l'hypothèse H_0 (équation 2.30) et pour l'hypothèse H_1 (équation 2.31) ainsi que le score BIC (équation 2.32) :

$$\mathrm{BIC}(\lambda_0) = \log L(x|\lambda_0) - m\frac{1}{2} n \log K \tag{2.30}$$

$$\mathrm{BIC}(\lambda_1, \lambda_2) = \log L(x_1|\lambda_1) + \log L(x_2|\lambda_2) - m\frac{1}{2}(2n)\log K \tag{2.31}$$

$$\Delta\mathrm{BIC} = \mathrm{BIC}(\lambda_1, \lambda_2) - \mathrm{BIC}(\lambda_0) \tag{2.32}$$

où m est une pénalité (théoriquement fixée à 1), n est le nombre de paramètres indépendants du modèle et K représente la taille de la fenêtre en vecteurs (Le et coll., 2007). Comme dans l'étape précédente, le signe du score indique l'hypothèse la plus vraisemblable. À l'issue de cette étape, les segments sont moins nombreux et plus longs.

Regroupement hiérarchique

La troisième étape consiste à fusionner les étiquettes des segments non contigus. À la différence des deux étapes précédentes, toutes les paires de segments sont considérées pour la fusion.

L'algorithme est ascendant et itératif. Il débute avec l'ensemble des segments résultant de l'étape précédente. À chaque tour, l'algorithme considère toutes les paires possibles, composées de deux segments différents. Pour chaque paire, les deux hypothèses H_0 et H_1 sont évaluées par l'estimation des modèles gaussiens sur les segments qui la composent puis les critères BIC et le score BIC sont calculés. La paire sélectionnée pour la fusion est celle qui présente le score BIC minimal négatif. Lorsqu'une paire est fusionnée, une étiquette commune est attribuée aux deux segments qui la composent, de sorte que tous les vecteurs regroupés sous cette étiquette seront employés par la suite pour l'estimation du modèle de la paire. L'algorithme s'arrête lorsque le score BIC minimal devient positif.

2.4.3.2 Algorithme de regroupement probabiliste EM

L'objectif de l'algorithme est de déterminer le groupe d'appartenance le plus probable pour chaque instance du corpus en faisant l'hypothèse d'une loi de probabilité pour chaque groupe, dont les paramètres sont également à déterminer (Witten et coll. 2005). Pour cela, la méthode est itérative et applique l'algorithme EM présenté précédemment. Nous présentons d'abord le fonctionnement de l'algorithme EM pour le regroupement. Nous verrons ensuite comment la méthode s'applique et, en particulier, comment elle permet de déterminer le nombre de groupes de manière non-supervisée.

Regroupement pour un nombre de groupes fixe

L'objectif est d'évaluer un GMM pour déterminer les groupes de vecteurs. Chaque groupe est représenté par une gaussienne du mélange. Nous considérons que le nombre de groupes est fixe. Après une initialisation des paramètres des gaussiennes, les deux étapes de l'algorithme EM sont appliquées de manière itérative.

Dans la première étape dite d'espérance (*expectation* en anglais), l'algorithme de regroupement EM estime la probabilité d'appartenance de chaque vecteur à chaque groupe en considérant les valeurs courantes des paramètres des modèles (estimées lors de l'itération précédente). Si l'on considère la probabilité d'appartenance $p(A|x)$ du vecteur x au groupe A, celle-ci peut s'exprimer par le principe de Bayes de la manière suivante :

$$p(A|x) = \frac{p(x|A)\,p(A)}{p(x)} = \frac{1}{\sqrt{2\pi}\sigma_A} e^{-\frac{(x-\mu_A)^2}{2\sigma^2}} \frac{p(A)}{p(x)} \tag{2.33}$$

La quantité $p(x|A)$ exprime la probabilité d'émission du vecteur x par le groupe A, représentée par une gaussienne de paramètres (μ_A, σ_A). La probabilité $p(A)$ peut être estimée par le ratio du nombre d'éléments affectés au groupe A lors de l'itération précédente. La probabilité $p(x)$ est partagée par les probabilités d'appartenance aux autres groupes. Il est possible de l'éliminer en sommant les probabilités d'appartenance de l'instance à tous les

groupes et en normalisant chaque probabilité individuelle par cette somme. La probabilité est aussi calculée pour les autres groupes.

Les nouvelles valeurs des probabilités d'appartenance permettent la réévaluation des paramètres des modèles de groupes ; c'est l'étape de maximisation. Pour le calcul des paramètres, les vecteurs x_i et leur probabilités d'appartenance $p(A|x_i)$ (dans la cas du groupe A) sont employés :

$$\mu_A = \frac{\sum_{i=1}^{N} p(A|x_i)x_i}{\sum_{i=1}^{N} p(A|x_i)} \qquad \sigma_A = \frac{\sum_{i=1}^{N} p(A|x_i)\left(x_i - \mu_A\right)^2}{\sum_{i=1}^{N} p(A|x_i)} \tag{2.34}$$

Enfin, on calcule la vraisemblance du GMM sur les données pour quantifier la "pertinence" du modèle aux données. Le calcul considère pour chaque vecteur la vraisemblance d'être généré par le GMM. Cela consiste à calculer la somme pondérée des probabilités *a posteriori* $p(x_i|k)$ d'un vecteur x_i connaissant un modèle k. Les poids sont les probabilités de chaque modèle, représentés par le ratio du nombre de vecteurs associés au groupe k. Ensuite, on multiplie les vraisemblances calculées pour chaque vecteur pour obtenir la vraisemblance globale, exprimée par la quantité $\prod_i \left(\sum_k p_k p(x_i|k) \right)$ où p_k représente la probabilité du groupe k.

Application de la méthode et détermination du nombre de groupes

D'abord, la méthode applique l'algorithme suivant une validation croisée à 10 sous-ensembles. L'estimation du GMM est effectué sur neuf des dix sous-ensembles et la log-vraisemblance du modèle est calculée sur le dixième sous-ensemble. L'opération d'estimation du GMM et d'évaluation de la vraisemblance est répétée 10 fois, en changeant le sous-ensemble laissé de côté. Les dix valeurs de vraisemblance sont moyennées.

Par ailleurs, la méthode permet aussi de déterminer le nombre de groupes. Pour cela, une phase supplémentaire est ajoutée, elle aussi itérative. La méthode commence avec un seul groupe. À chaque itération, le GMM est évalué et testé en validation croisée à 10 sous-ensembles et la vraisemblance moyenne est comparée à celle de l'itération précédente. Si la variation est supérieure à un seuil fixé, alors on incrémente le nombre de groupes et on recommence la phase d'estimation du GMM. Sinon, on garde les groupes de l'itération courante.

2.5 Bilan de l'état de l'art

L'étude décrite dans le chapitre a permis d'aborder plusieurs problèmes de la thèse. Pour éclaircir la compréhension du concept de scène, nous avons d'abord étudié la notion de contexte. Nous avons retenu la décomposition de celui-ci en plusieurs éléments dont l'environnement, les personnes et les objets. Concernant le problème de reconnaissance de scène, nous avons mis en évidence un processus composé de trois étapes principales : la mesure, la

perception et la décision. C'est la seconde étape qui est la plus intéressante car elle effectue la transformation des mesures de capteurs en symboles abstraits qui peuvent être interprétés.

Le problème de la reconnaissance de scène a aussi été abordé par la description de tâches associées. Ainsi, les tâches de localisation, de reconnaissance de l'ambiance sonore de l'environnement ou de reconnaissance de l'activité physique sont pertinentes pour la reconnaissance de scène sur smartphone et les travaux associés ont indiqué des résultats satisfaisants.

Plusieurs problèmes liés au smartphone ont aussi donné lieu à une réflexion dans ce chapitre. D'abord, le contexte du smartphone a été mis en évidence, notamment par la position et la localisation de celui-ci. Plusieurs stratégies pour l'éviter ou pour le reconnaître ont été proposées. Nous avons aussi évoqué la diversité des sources de données dans les travaux rapportés. L'accéléromètre et le microphone semblent être les sources les plus pertinentes, si l'on considère notre état de l'art et la comparaison proposée dans le chapitre.

Enfin, le chapitre a été l'occasion d'étudier les méthodes mises en place pour la réalisation de systèmes et d'expérimentations dans différents travaux. Nous retenons des descripteurs de données pour les sources et des classifieurs populaires, aux bons résultats, tels que les arbres de décision C4.5 ou forêt d'arbres décisionnels, le réseau bayésien ou le mélange de gaussiennes (GMM). Nous avons aussi décrit les réseaux de neurones et, en particuliers, ceux dits "profonds" pour leur ascendance récente.

Dans le chapitre suivant, nous abordons un autre problème évoqué dans l'introduction : celui de la constitution d'un corpus de données. Comme nous allons le voir, il s'agit d'une tâche complexe, qui soulève de nombreuses questions, en particulier dans le cadre de l'utilisation d'un smartphone.

CHAPITRE 3

Collecte de données et définition des contextes

Comme nous l'avons dit en introduction du manuscrit, la connaissance des scènes est floue. De plus, notre approche pour le problème de reconnaissance de scène repose sur une méthode d'apprentissage automatique supervisé qui nécessite des données. La constitution d'un corpus est donc justifiée, d'une part, pour l'observation des scènes en vue d'améliorer notre connaissance et, d'autre part, pour l'évaluation des algorithmes de reconnaissance.

Dans un premier temps, le chapitre définit les caractéristiques des données à collecter. Les scènes d'intérêt *a priori* y sont décrites, les sources énumérées et des caractéristiques plus pragmatiques sur les données sont définies (telle que la synchronisation des différentes sources).

Nous présentons ensuite le protocole pour la réalisation d'une collecte. Celui-ci est conçu pour enregistrer des données de différentes sources, synchronisées, annotées, pendant de longues durées et dans des conditions réelles. L'annotation d'une collecte *in vivo* soulève des questions, que nous identifions et pour lesquelles nous fournissons une solution. La sécurité des donnés pendant l'enregistrement et le respect de la vie privée de l'utilisateur représentent une autre problématique, à laquelle nous répondons aussi. Nos différentes solutions permettent de définir le protocole de collecte général.

Ensuite, l'application développée pour la collecte et appelée RECORDME est présentée. Sa conception est influencée par les caractéristiques des données et du protocole. L'application a été évaluée par des mesures de performance objectifs et des avis d'utilisateurs.

Enfin, les deux collectes effectuées avec l'application sont détaillées dans une dernière section. La collecte de scènes est la plus importante des deux et a permis d'enregistrer plus de 570 heures de données. La seconde collecte complète la première par l'enregistrement d'éléments de scènes qui n'avaient pas pu être annotés dans la première collecte (telles que les activités physiques simples et la position du smartphone).

3.1 Définition des caractéristiques des données

Pour l'acquisition du corpus, il est d'abord nécessaire de définir les données d'intérêt et le contexte de leur capture.

TABLE 3.1: Scènes d'intérêt pour la collecte, issues de l'état de l'art

Exemples de l'état de l'art	**Scènes considérées**
Salon, cuisine, salle de bains (Peltonen et coll. (2002)) Préparer le repas, passer l'aspirateur (Dernbach et coll. (2012))	Situation du domicile
Travail sur ordinateur (Bao et Intille (2004))	Activité de travail dans un bureau
Rue et parc (Défi D-CASE)	Situation en extérieur
Supermarché et restaurant (défi D-CASE)	Lieu de vie public
Bus et métro (défi D-CASE), déplacements motorisés (Reddy et coll. (2010))	Déplacement en transport motorisé
Conversation (Kern et coll. (2007))	Conversation
Marche, montée ou descente d'escaliers, vélo (Bao et Intille (2004)) Activités immobiles en posture assise, debout ou couché (Bao et Intille (2004), Kwapisz et coll. (2011))	Activités physiques simples

3.1.1 Définition des scènes d'intérêt

À partir des travaux de reconnaissance issus de l'état de l'art, nous avons sélectionné un ensemble de scènes que nous présentons dans la table 3.1. Les exemples de scènes sont présentés à gauche et, à droite, nous indiquons notre représentation. Les exemples du domicile sont nombreux et précis. Ils décrivent des lieux de vie et des activités effectuées. Nous choisissons une représentation plus générale pour la collecte en considérant globalement les situations du domicile. L'activité de travail au bureau est illustrée par le travail sur ordinateur. Nous suggérons aussi les réunions de travail et les moments de pause comme scènes pertinentes au travail. Les lieux en extérieur sont évoqués, notamment par les lieux de rue et de parc en milieu urbain. Les lieux de vie dits public sont aussi considérés comme des scènes d'intérêt pour l'application industrielle, avec le supermarché et le restaurant comme exemples. Nous en suggérons d'autres comme les lieux de spectacle (théâtre, cinéma, salle de concert), les administrations ou les gares de transport. La table mentionne aussi les transports motorisés car ils occupent une part importante de la vie quotidienne des personnes et peuvent, à ce titre, être pertinentes pour l'application industrielle. Le bus et le métro sont cités comme exemples, mais nous considérons aussi la voiture, le train, le tramway, l'avion ou encore le bateau.

Les lignes suivantes de la table se caractérisent par l'absence d'association à un lieu. Il s'agit de situations particulières, créées par une action ou une activité. Ainsi, nous nous intéressons aux situations de conversation, aux activités physiques simples de marche, de montée ou descente d'escaliers ou de vélo et aux activités immobiles qui sont nombreuses dans le mode de vie sédentaire actuel.

3.1.2 Définition des sources d'intérêt

Dans l'état de l'art, nous avons évalué plusieurs sources de données suivant leur pertinence dans les travaux de reconnaissance et suivant des critères pragmatiques d'utilisation. La table 3.2 résume l'ensemble des sources considérées dans la thèse.

La revue des travaux de reconnaissance a montré la pertinence de l'accéléromètre, du microphone et des sources de positionnement d'antennes-relais et du GPS, c'est pourquoi elles sont toutes présentes dans la table. Nous avons également évoqué l'usage du Wi-Fi, du Bluetooth, du gyroscope et du magnétomètre en les présentant comme des sources complé-

Catégorie	Sources
Inertielle	Accéléromètre, gyroscope
Localisation	GPS, antennes-relais, Bluetooth, Wi-Fi
Capteur d'ambiance	Magnétomètre, luminosité, proximité, baromètre, microphone
Journal de communications	Appels, sms, données
Journal d'actions	Écouteurs, batterie, écran, applications

TABLE 3.2: Tableau des sources de données

mentaires ou avec un usage limité pour une tâche particulière. Nous souhaitons les prendre en compte car elles peuvent être pertinentes pour d'autres tâches de reconnaissance. Enfin, certaines sources absentes des travaux mentionnés sont présentes sur les smartphones [1]. C'est le cas de capteurs ambiants pour la luminosité et la pression atmosphérique. Également, le journal du fonctionnement de l'appareil fournit des informations sur l'allumage de l'écran, l'usage des applications ou encore les communications passées. Nous suggérons que la collecte de ces éléments peut ensuite permettre l'interprétation ou l'identification des scènes, par l'étude de corrélations par exemple.

3.1.3 Définition des caractéristiques des données

La caractérisation des sources et des scènes d'intérêt, combinée au bilan des travaux de l'état de l'art laisse imaginer les caractéristiques des données souhaitées. Nous les formulons précisément dans cette section. Après cela, nous recherchons l'existence d'un corpus de données correspondant et, à défaut, l'existence d'outils de collecte pertinents.

3.1.3.1 Caractéristiques des données

Nous souhaitons obtenir des données multimodales et synchronisées ; réelles et annotées ; et collectées de manière continue. La synchronisation est justifiée par la présence de plusieurs sources. Ensuite, nous souhaitons acquérir des données réelles, collectées *in vivo*, car les scènes d'intérêt représentent des situations quotidiennes, difficiles à reproduire en laboratoire (c'est le cas des transports motorisés par exemple). En outre, l'appareil visé par l'application industrielle est le smartphone. La collecte de données directement depuis l'appareil permet de travailler avec des données représentatives du contexte de fonctionnement du système final de reconnaissance. En outre, l'usage du smartphone pour la collecte est aussi opportun car l'appareil est emporté dans de nombreux endroits par son utilisateur. L'annotation des données est nécessaire pour pouvoir analyser le corpus acquis et, par suite, procéder à une phase d'apprentissage supervisé pour construire le système de reconnaissance de scène. Enfin, l'enregistrement continu des données est important pour la capture des transitions entre les scènes.

1. La documentation en ligne des développeurs pour Android répertorie de nombreuses sources accessibles. L'adresse suivante permet d'accéder à la page d'entrée du sujet http://developer.android.com/guide/topics/sensors/sensors_overview.html.

3.1.3.2 État de l'art des bases et outils de collecte de smartphone

Comme nous l'avons dit dans le chapitre de l'état de l'art, le smartphone a fait l'objet de nombreuses recherches. Il existe ainsi un petit nombre de corpus accessibles. Parmi les premières bases collectées, les travaux de Pentland et coll. (2006), au milieu des années 2000, ont lancé le projet *Reality Mining* pour la collecte de données sur téléphones. Les smartphones démarraient à peine, aussi les téléphones employés disposaient de peu de capteurs. Le corpus collecté se compose des informations d'appels passés, des connexions d'appareils en Bluetooth, des connexions aux antennes-relais, de l'usage des applications de l'appareil et d'informations sur l'état de fonctionnement telles que le niveau de charge de la batterie. Les annotations ont été effectuées semi-automatiquement : le système identifiait des antennes-relais préalablement associées par l'utilisateur à des étiquettes de lieux sémantiques tels que le domicile ou le lieu de travail. 90 volontaires ont participé à cette collecte. Les limites de sources et de labels de cette base justifient que nous ne l'ayons pas exploité.

En 2010, Kiukkonen et coll. (2010) ont réalisé une collecte à grande échelle dans la région de Lausanne, en partenariat avec Nokia et connue sous le nom de *Lausanne Data Collection Campaign* (abrégé par LDCC). La base de données se compose de coordonnées de localisation par GPS, d'informations sur l'usage des applications (prise de photo ou vidéo, agenda, répertoire téléphonique, communications par SMS et appels) et de données d'accélérations et acoustiques. Près de 200 volontaires y ont participé. Cependant, nous n'avons pas utilisé ce corpus pour plusieurs raisons : l'enregistrement des échantillons sonores ne convenait pas (effectué par période de 30 secondes, espacées de plusieurs minutes) ; le corpus est devenu inaccessible après la période d'évaluation ; les annotations associées aux données ne correspondent pas aux scènes que nous recherchons.

Récemment, Wagner et coll. (2014) de l'université de Cambridge ont rapporté la réalisation d'une collecte d'une très grande ampleur intitulée *Device Analyzer*. Les auteurs se targuent de plus de 26000 téléchargements de leur application de collecte et rapportent 16000 participants volontaires (dont plus de 4700 auraient participé pendant plus d'un mois). De nombreuses sources sont collectées, bénéficiant des divers capteurs embarqués sur les smartphones employés et des possibilités d'accès à de nombreuses informations de fonctionnement de l'appareil (dont nous citons quelques exemples : l'usage d'une carte externe de mémoire, le nombre de contacts dans le répertoire, la taille de l'espace mémoire libre ou la version du système d'exploitation). Cependant, l'absence de données sonores et d'annotations correspondant aux scènes recherchées ne nous ont pas incités à exploité ce corpus.

Les caractéristiques des données que nous avons énoncées deviennent des contraintes pour la sélection d'une base de données satisfaisante. Aussi, nous envisageons d'effectuer notre propre collecte et, pour cela, nous rapportons les résultats d'une recherche d'outils adaptés.

La recherche s'est principalement orientée vers les outils sur smartphones car ce dernier est l'appareil visé par l'application du partenaire industriel et il a été employé dans les

trois collectes rapportées. Il paraît également très adapté à l'enregistrement des données suivant les caractéristiques énoncées. Ses nombreuses sources embarquées ont déjà été évoquées et l'accès aux indications d'horodatage fournies avec les données permet la synchronisation. Il est idéal pour une collecte *in vivo* car la plupart des personnes qui en possèdent un l'emportent dans de nombreuses situations quotidiennes. Grâce à ses ressources, il est envisageable d'effectuer des enregistrements continus sur plusieurs heures, garantissant la capture de transitions entre des scènes. Cependant, le problème de l'annotation reste ouvert pour le moment.

L'outil *Open Data Kit* (abrégé ODK) est conçu pour la collecte de données sur smartphone sans requérir des expérimentateurs des compétences en programmation informatique. À son lancement en 2010, l'outil prévoit trois fonctions (Hartung et coll. 2010) : une interface graphique pour renseigner les caractéristiques de la collecte (par exemple les sources, l'échantillonnage, ou la durée) ; la création automatique de l'application et l'installation sur les appareils pour la collecte ; le transfert et le stockage des données sur des serveurs dédiés. Au lancement, peu de sources de données étaient disponibles (appareil photo, GPS et usage de l'écran tactile). Une mise à jour a été proposée en 2013 (Brunette et coll. 2013) incluant notamment la possibilité de gérer plus de sources de données (internes et externes *via* des liaisons Bluetooth et USB). Cependant, nous n'avons pas pu déterminer précisément le niveau de sécurité et d'anonymat des données prévu par l'outil. Pour cette raison, nous ne l'avons pas utilisé pour notre collecte.

Le résultat le plus adapté au cours de nos recherches est le projet Funf[2] lancé en 2011 et publié par Aharony et coll. (2011) qui permet la collecte et la sauvegarde sur un serveur de nombreuses sources du smartphone. Il se présente sous trois formes qui reposent sur le même code : une application, un service et une bibliothèque de code libre d'accès. L'application, appelée Funf Journal[3] et disponible sur la plateforme de téléchargement Google Play, permet la collecte de nombreuses sources et le transfert vers un serveur. Mais la configuration de ces sources est limitée et ne permet pas un enregistrement continu pendant une longue période. Le service, appelé *Funf in a Box*, est destiné à des expérimentateurs qui souhaitent disposer d'un outil sans avoir à le développer. Il permet de créer et configurer l'application à travers un formulaire, puis de la télécharger et, enfin, de disposer d'un compte Dropbox pour sauvegarder les données des participants. L'inconvénient de cette solution est la dépendance à un service de sauvegarde extérieur et les problèmes de respect de la vie privée que cela pose. Enfin, la bibliothèque propose le code sous licence LGPL[4]. L'étude du code confirme la présence de toutes les sources qui nous intéressent. Cependant, son usage a semblé complexe et peu documenté et ne disposait pas d'exemples d'intégration à une application Android. À l'inverse, la documentation en ligne d'Android présente de nombreux

2. http://www.funf.org/
3. https://play.google.com/store/apps/details?id=edu.mit.media.funf.journal
4. *GNU Lesser General Public* ou licence publique générale GNU est une licence pour gérer l'utilisation de logiciels. Elle est moins restrictive que la licence GNU GPL, dont elle découle, car elle permet de créer un outil ou une ressource dite *propriétaire* qui exploite des outils libres, eux-mêmes sous licence LGPL.

exemples sur l'usage des capteurs et les objets nécessaires à leur manipulation. C'est pourquoi nous n'avons pas choisi d'utiliser la bibliothèque du projet Funf mais avons décidé de créer notre propre application pour la collecte.

3.2 Les problématiques de la collecte

La réalisation d'une collecte de données variées et annotées, par des volontaires, dans des situations réelles, soulève deux questions. La première porte sur la méthode d'annotation, qui doit fournir des étiquettes de qualité, en limitant l'influence sur la collecte *in vivo* et sur la qualité des données collectées. La seconde concerne la sécurité des données enregistrées et le respect de la vie privée des participants, aussi bien pendant la collecte que lors des traitements ultérieurs. L'ensemble des solutions apportées à ces questions permet de définir le protocole général de la collecte.

3.2.1 Le choix de l'auto-annotation et les problématiques associées

Nous avons envisagé plusieurs solutions pour l'annotation des données. La présence d'un expérimentateur pour cette tâche a été rapidement écartée. En effet, soit la collecte est *in vivo* et l'expérimentateur suit le volontaire, ce qui est trop contraignant pour être mis en place ; soit les scènes sont reproduites en laboratoire et leur réalisation est supervisée, ce qui n'est pas envisageable pour certaines scènes telles que les transports.

L'auto-annotation apparaît comme le meilleur compromis pour obtenir des données réelles et annotées. Cependant, ce choix soulève d'autres questions. D'abord, puisque l'expérimentateur est absent, la collecte est non-supervisée et le volontaire est responsable de l'annotation. En particulier, celui-ci doit connaître les concepts à annoter et savoir comment les annoter. Cela nécessite l'usage d'un outil adapté et maîtrisé par le volontaire. De plus, ce dernier doit savoir quels concepts sont recherchés et comment les identifier pour les annoter.

Les solutions apportées à ces questions sont un moyen de mettre le volontaire dans de bonnes conditions pour l'annotation. Mais elles ne garantissent pas la validité des annotations récupérées. Cela introduit la seconde problématique de l'auto-annotation : comment évaluer les annotations des volontaires, notamment les descriptions des scènes et les horodatages ?

3.2.2 La sensibilisation du volontaire et l'outil d'annotation

Pour faciliter l'auto-annotation, notre solution repose sur l'information aux volontaires sur différents aspects de la collecte et des annotations. De plus, l'outil d'annotation doit être pensé pour favoriser la saisie.

3.2.2.1 L'information donnée au volontaire

La première étape pour l'auto-annotation est une phase d'information aux volontaires avant de démarrer la collecte. Elle permet de les sensibiliser aux différents éléments de scènes de la table 3.1 intéressants pour la collecte et aux règles à suivre pour l'annotation. Ces règles sont représentées dans la liste suivante.

1. Une scène est définie par un environnement et une activité réalisée.
2. Une scène change lorsque l'on change d'environnement ou d'activité.
3. L'annotation d'une scène doit être faite au début de celle-ci (à l'entrée dans un nouvel environnement ou lorsque l'on démarre une nouvelle activité) ; ne pas anticiper un changement en réalisant une annotation avant la scène.
4. Les transitions peuvent être déroutantes (par exemple, je suis dans la cage d'escalier de mon immeuble, suis-je déjà au "Domicile" ?) ; pas d'inquiétude, nous les prendrons en compte lors de la validation des annotations).
5. Ne pas changer ses habitudes et vivre sa vie normalement.
6. L'enregistrement est libre : l'application permet de démarrer et d'arrêter l'enregistrement de capteurs individuellement ou en groupe.

Les règles définissent le concept de scènes que nous considérons pour la collecte. Elles indiquent également le comportement à avoir lors d'un changement de scènes. La règle 3 précise que l'annotation doit être faite au début de la nouvelle scène. En outre, les transitions ne sont pas forcément nettes et il peut arriver des situations ambiguës comme la situation dans la cage d'escalier mentionnée dans la règle 4. Dans ce cas, l'interprétation est subjective et nous expliquons aux volontaires que ces transitions sont traitées précautionneusement lors du traitement.

3.2.2.2 L'outil d'annotation

Nous avons fait le choix d'intégrer l'outil d'annotation au smartphone, ce qui permet le contrôle de la collecte et des annotations à partir d'un seul appareil. L'outil d'annotation se présente sous la forme d'une interface, intégrée à l'application de collecte de données.

Nous avons pensé l'interface d'annotation comme un moyen efficace pour le volontaire de renseigner les annotations. Ainsi, l'interface se compose d'une page unique sur laquelle le volontaire dispose de plusieurs modes pour renseigner les descriptions : des menus déroulants, des champs de texte libre et des boutons radio. Les menus déroulant sont constitués des labels prédéfinis. Leur but est de faciliter l'annotation, en particulier pour les scènes les plus fréquentes. Les champs de texte libre sont présents pour permettre une description plus précise ou indiquer une scène absente des menus déroulants. Le changement d'un mode à l'autre se fait par simple pression d'un bouton.

Les quatre captures d'écran de la figure 3.1 illustrent la réalisation de cette interface. L'organisation de la page se décompose en trois parties, que l'on peut identifier sur les captures d'écran du haut. La première partie concerne la localisation du volontaire et invite à décrire

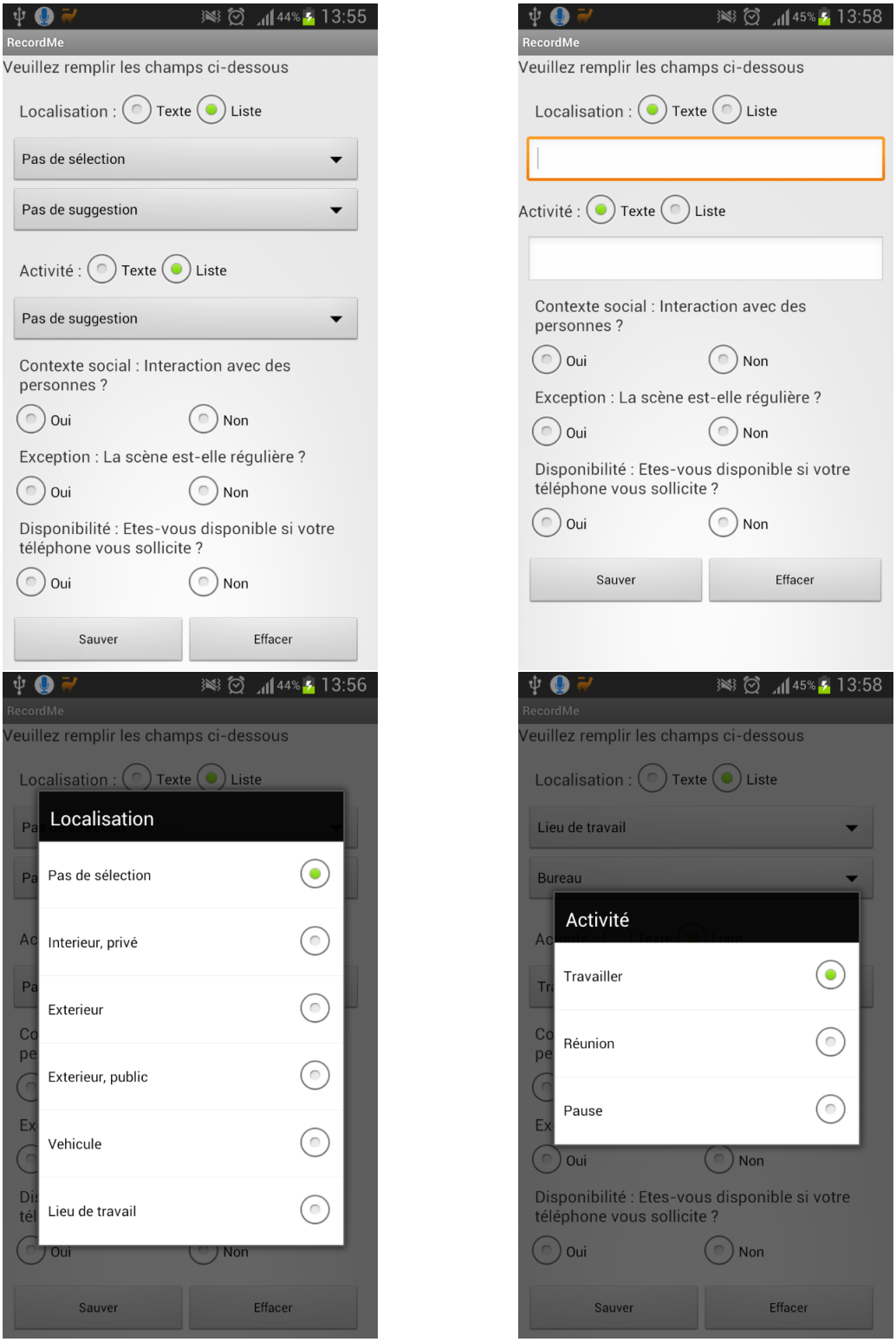


FIGURE 3.1: Illustrations de l'interface d'annotation de RECORDME dans plusieurs cas avec de haut en bas et de gauche à droite : mode "menu déroulant" et mode "texte libre" ; suggestion de lieux en bas à gauche et suggestion d'activités en bas à droite

le lieu. La seconde partie concerne la description de l'activité. La troisième partie, composée des trois rangées de boutons radio, concerne la présence de conversations et des infor-

mations annexes (la fréquence de la scène et la disponibilité du volontaire à répondre aux sollicitations de son téléphone dans la scène courante).

L'intérêt des menus déroulants réside également dans l'adaptation de leur contenu aux choix du volontaire. En effet, les environnements sont regroupés suivant des catégories, accessibles dans le premier menu de la capture en haut à gauche de la figure 3.1. La sélection d'une catégorie met à jour le second menu qui propose une liste d'environnements adaptés. En outre, la liste des activités possibles est également mise à jour suivant la catégorie de l'environnement. Les captures d'écran du bas de la figure 3.1 illustrent les catégories d'environnement et d'activité.

3.2.3 L'évaluation des annotations

Le cadre favorable créé par la phase d'information et la facilité d'utilisation de l'outil ne garantissent pas la qualité des annotations. C'est pourquoi la seconde partie de notre solution consiste à évaluer les annotations rapportées. Pour cela, nous formulons plusieurs hypothèses et tirons profit de données objectives.

3.2.3.1 Les hypothèses pour l'évaluation

L'évaluation repose d'abord sur une observation particulière. Les scènes fréquentes sont associées à des lieux particuliers : le travail en bureau s'effectue dans un bâtiment identifié, le domicile est souvent unique et identifié, les sorties telles que les courses ou les loisirs s'effectuent dans des lieux dédiés. Par conséquent, on peut imaginer que la quantité de lieux à vérifier est faible.

De plus, beaucoup de transitions s'effectuent par changement de lieu et beaucoup nécessitent la sortie ou l'entrée dans un bâtiment. Nous avons relevé dans l'état de l'art l'article de Marmasse et Schmandt (2000) qui indiquent la possibilité de détecter les entrées et sorties dans des bâtiments à partir des traces de coordonnées fournies par le GPS. Dans notre cas, les coordonnées ne suffisent pas, il faut pouvoir les relier à un contenu sémantique. Nous avons également mentionné dans l'état de l'art le projet OpenStreetMap (2014) dont le but est de proposer une carte géographique enrichie de nombreux labels sémantiques. L'exploitation combinée des coordonnées et de la carte offre un moyen pertinent d'évaluer les transitions et, par suite, les scènes.

D'autres données peuvent fournir des informations objectives pour l'évaluation des annotations. La connexion à une borne Wi-Fi indique que la proximité à celle-ci est inférieure à quelques dizaines de mètres. En outre, la portée de la borne au sein d'un bâtiment permet de compléter les informations de localisation du GPS. Une autre information peut provenir du branchement du téléphone pour recharger la batterie. Si l'on suppose que l'utilisateur reste à proximité de son téléphone pendant la durée du chargement, on peut estimer une période d'immobilité de l'utilisateur. Encore une fois, cette information complète celles fournies par le GPS puisque la plupart des prises électriques sont dans des bâtiments.

Plusieurs solutions objectives apparaissent pour l'évaluation de la localisation. Cependant, celles-ci ne permettent pas l'évaluation de l'activité de la personne. Dans ce cas, nous devons considérer que la description de l'utilisateur est correcte. Les descriptions des activités proposées pour la collecte sont spécifiques à certains lieux et assez générales (par exemple au lieu de travail, il est possible de travailler à son bureau, d'être en réunion ou en pause).

3.2.3.2 La procédure d'évaluation

Le traitement de l'évaluation tire profit des données objectives collectées. En particulier, il s'agit d'identifier les périodes où le volontaire peut être localisé grâce à la présence de coordonnées géographiques du GPS dans les données. Par suite, les traces de coordonnées sont traitées pour déterminer les étiquettes sémantiques des lieux visités. Les indications temporelles complètent la description des étiquettes. Le traitement est complété par la détection de zones d'immobilité à partir des connexions aux bornes Wi-Fi et des périodes de rechargement de la batterie. Le résultat correspond à une séquence de lieux obtenus par le traitement des données objectives qui peut être comparé au fichier d'annotations fourni par l'utilisateur.

En pratique, ce traitement n'a pas pu être automatisé entièrement, en particulier car il existe de nombreuses zones vides d'étiquettes dans la carte enrichie d'OpenStreetMap. C'est pourquoi, dans de nombreux cas, il a fallu visualiser la trajectoire suivie sur une carte pour identifier la séquence des lieux. Pour cela, nous avons employé l'outil MyTourbook [5] qui permet l'affichage de trajectoires sur la carte d'OpenStreetMap avec les indications temporelles et éventuellement d'autres informations telles que l'altitude ou la vitesse.

Nous représentons une capture d'écran de l'outil dans la figure 3.2 sur laquelle on distingue la carte géographique et la trajectoire effectuée. Des agrandissements sont possibles sur la carte, ce qui permet d'afficher certaines étiquettes associées à des lieux. Le graphique du bas de la figure affiche la vitesse suivant l'axe du temps. Les points de la trajectoire sont synchronisés avec ceux du graphique, si bien qu'il est possible de visualiser la vitesse pour n'importe quel point de la trajectoire.

Malgré cela, des zones d'incertitude peuvent subsister. D'abord, les entrées et sorties de bâtiments ne sont pas toujours nettement observables par les coordonnées fournies. Cela est dû à des phénomènes physiques de déviation des ondes électromagnétiques du signal du GPS par les bâtiments, ce qui peut créer des interférences aux moments de ces transitions. L'incertitude existe également dans l'interprétation de certaines transitions, comme nous l'avons déjà remarqué, notamment avec l'exemple de la cage d'escaliers dans la section 3.2.2.1. Par conséquent, nous associons une étiquette d'incertitude à ces transitions dont la durée est réduite au minimum (deux minutes en moyenne sur les enregistrements traités). Certains lieux invérifiables sont également associés à une étiquette d'incertitude.

5. Accessible en téléchargement à l'adresse http://mytourbook.sourceforge.net/mytourbook/. Le public visé par cet outil regroupe les pratiquants de sport de nature (cyclisme, randonnée, course à pied, sport de glisse, etc) qui souhaitent représenter voire analyser leurs sorties.

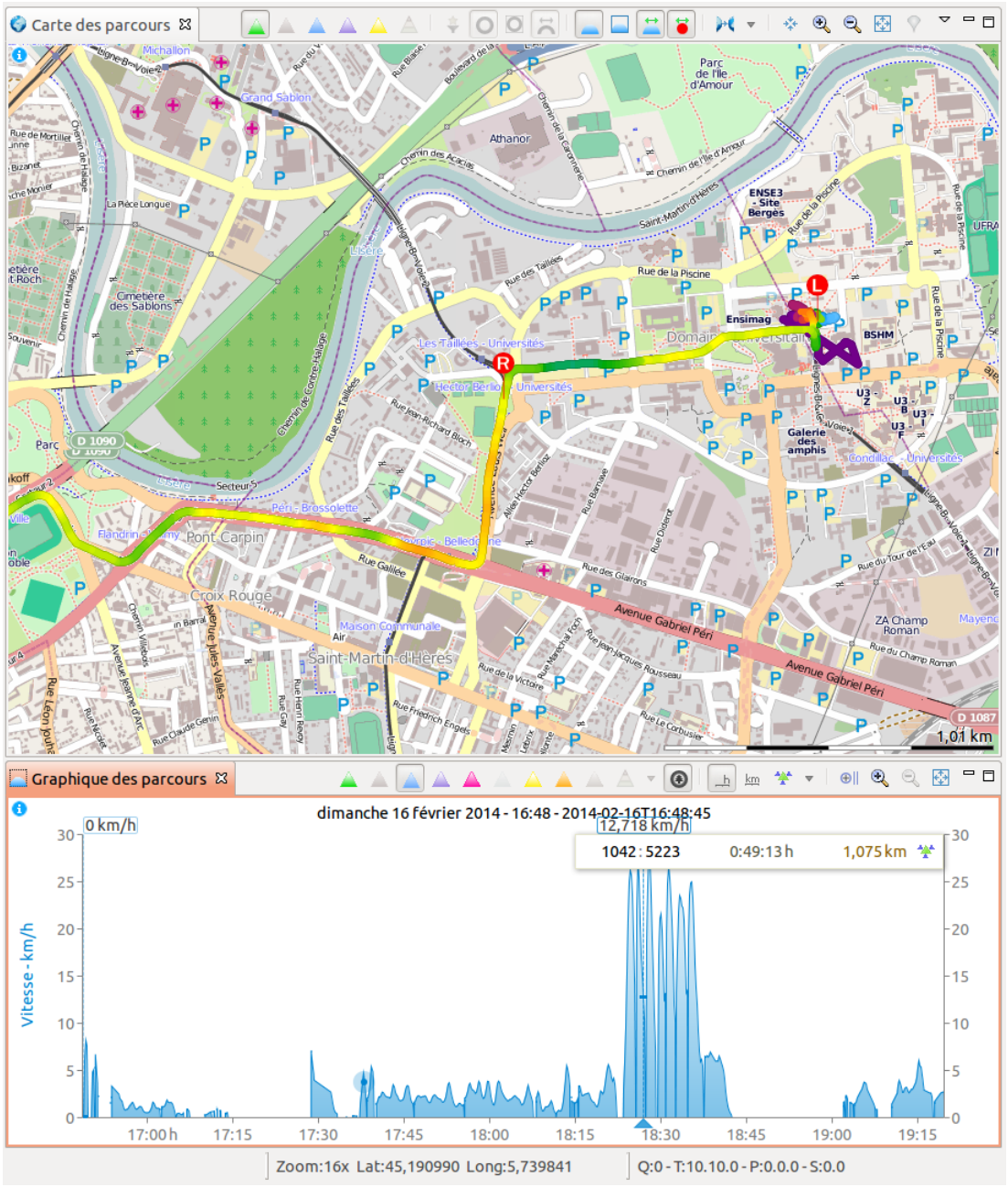


FIGURE 3.2: Capture d'écran de MyTourbook

C'est le cas lorsque le signal GPS est temporairement ou totalement absent et que les autres sources dites objectives ne permettent pas l'identification.

Lorsque ces zones subsistent, il est prévu d'en discuter avec le volontaire en lui présentant les annotations corrigées ainsi que la visualisation de sa trajectoire afin d'essayer d'identifier la scène.

Le recours à la visualisation des trajectoires des volontaires peut poser des problèmes liés au respect de la vie privée. La section suivante est dédiée à cette problématique ainsi qu'à la sécurité des données.

3.2.4 Les problématiques de sécurité et du respect de la vie privée

La seconde problématique de la collecte porte sur la sécurité des données et le respect de la vie privée des volontaires qui participent. Nous présentons les risques encourus afin de mieux cibler les solutions à apporter.

3.2.4.1 L'identification des risques

Le choix d'effectuer les enregistrements sur des smartphones dans des situations quotidiennes réelles entraîne le risque de perte, vol ou dégradation de l'appareil. Dans ces conditions, les données collectées et présentes sur l'appareil peuvent être récupérées par de tierces personnes. Ainsi, il paraît important de déterminer le niveau de confidentialité de ces données pour pouvoir mieux les protéger.

Nous avons identifié les sources dont le contenu ne doit pas être divulgué :

— les échantillons sonores : le principal danger de la collecte du son est l'extraction du contenu des conversations ;

— les chaînes de caractères : les identifiants des appareils alentours (antennes-relais, bornes Bluetooth, bornes Wi-Fi), numéros de téléphone ou noms des applications utilisées sont autant d'éléments sensibles ;

— les coordonnées de localisation : les coordonnées géographiques fournies par le GPS du smartphone ou celles des antennes-relais ou bornes Wi-Fi sont des éléments très sensibles.

Au-delà de la période temporaire de stockage des données sur le smartphone, les risques concernent également le transfert des données et la sauvegarde sur le serveur distant pour la durée du traitement. Ainsi, les mêmes risques d'accès par de tierces personnes sont à considérer. Également, puisque le processus d'évaluation des annotations prévoit une phase où les coordonnées géographiques sont décryptées, les expérimentateurs ont accès au contenu des trajectoires suivies. Ainsi, il est important de protéger les volontaires contre la divulgation du contenu de leurs données et de leur identité, suivant le respect de leur vie privée.

3.2.4.2 Les solutions proposées

Nous évoquons les solutions proposées suivant l'ordre logique du parcours des données : du stockage sur le smartphone jusqu'à celui sur le serveur, en passant par le transfert. La première mesure proposée consiste à traiter à la volée les données des sources préalablement identifiées pour en éliminer le contenu brut et n'en garder qu'une représentation, qui ne permet pas d'en extraire le contenu jugé sensible :

— échantillons sonores : des descripteurs sont calculés à la volée et le signal brut n'est pas conservé ; ces descripteurs sont calculés sur des durées suffisamment longues pour ne pas permettre la reconstitution inverse du signal ou l'extraction du contenu de la parole ;

— chaînes de caractères : les chaînes sont combinées à un identifiant unique par utilisateur et ensuite cryptées au moyen d'une fonction de hachage et de l'algorithme SHA-256 [6] ;

— coordonnées de localisation : elles sont translatées avant d'être sauvegardées.

La deuxième mesure consiste à effacer les données du téléphone au-delà de 48 heures après la date de leur collecte.

Concernant le transfert des données, nous envisageons un mode filaire *via* une liaison USB et un mode sans fil *via* une connexion Internet. Le transfert filaire n'est effectué que vers une machine identifiée du laboratoire. Celle-ci est protégée par accès physique (l'accès au laboratoire est réglementé) et logique (l'utilisation est protégée par mot de passe). Le transfert sans fil est effectué au moyen d'une fonctionnalité intégrée à l'application de collecte. Celle-ci utilise le protocole SCP (*Secure Copy Protocol*, ou protocole de copie sécurisé) qui crypte les données et les transmet au serveur distant qui les transfère à son tour vers une machine du réseau interne du laboratoire, inaccessible depuis l'extérieur.

La sécurité des données sur le serveur de stockage est assurée par la double protection physique et logique des machines dans le laboratoire. Par ailleurs, après l'évaluation des annotations, les coordonnées géographiques sont à nouveau translatées et l'ensemble des données est sauvegardé anonymement, de sorte qu'on ne puisse plus identifier le volontaire qui les a produites.

3.2.5 Le protocole de collecte général

En guise de bilan de la section sur les problématiques de la collecte, nous présentons le protocole général pour la collecte de données. C'est également l'occasion de décrire les traitements qui n'ont pas été mentionnés tels que la mise en forme des fichiers transférés, la synchronisation des données des différentes sources et l'indexation des fichiers pour la sauvegarde. Le protocole peut être résumé par le schéma de la figure 3.3 dont les principales étapes sont les suivantes :

1. Accueil et information du participant : le volontaire est informé de la finalité de la collecte, des détails du traitement, des données collectées et, en particulier, des données sensibles et des mesures mises en place pour protéger ces données ; également, on installe l'application sur le smartphone du participant et on l'informe du fonctionnement de l'application (suivant le document en Annexe (page 165)), des scènes d'intérêt et du comportement attendu pour les annotations ;
2. Collecte des données : on note en particulier les mesures de cryptage et d'effacement des données qui sont effectuées pendant cette étape ;
3. transfert des données : il est effectué suivant les deux modes sécurisés présentés ;
4. post-traitement : l'étape consiste d'abord en la mise en forme des données ; des discontinuités temporaires dans les données collectées sont recherchées, causées par des

6. Algorithme de la famille SHA-2 qui retourne un résultat d'une longueur fixe de 256 bits.

dysfonctionnements éventuels de l'application ; la synchronisation des données suivant les sources est effectuée, grâce aux indications temporelles associées aux données, suivant l'hypothèse d'une horloge commune aux différentes sources ; un fichier d'indexation des données est créé, qui contient des informations générales sur les types de données disponibles, la durée de l'enregistrement, le chemin des fichiers ; c'est aussi l'étape d'évaluation des annotations ;

5. sauvegarde de longue durée : les données sont rendues anonymes pour être sauvegardées.

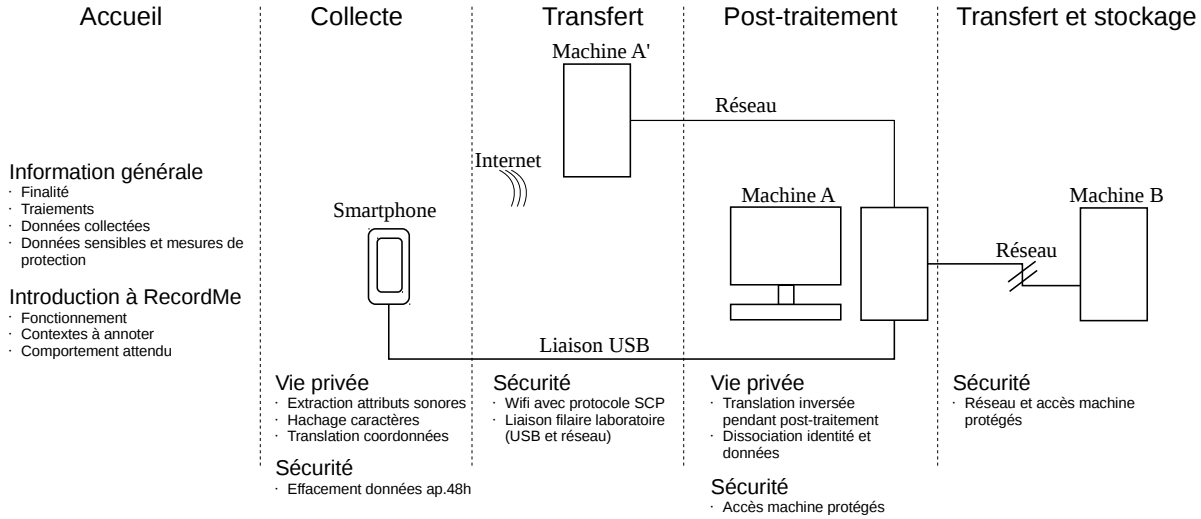


FIGURE 3.3: Schéma du protocole de collecte de données non-supervisées

Le protocole présenté est le résultat de la discussion avec des représentants locaux de la CNIL[7]. Il a été validé par leurs soins et est disponible suivant la référence 0750532. Nous rapportons en Annexe (page 174) le document qui a servi à sa validation.

3.3 L'application RECORDME

La section présente l'application RECORDME (Blachon et coll. (2014b)) conçue pour la collecte de nombreuses sources embarquées sur un smartphone. Les caractéristiques des données précédemment présentées ont fortement orienté le développement de l'application, en particulier pour permettre un enregistrement continu. L'interface graphique a également été pensée pour donner le contrôle à l'utilisateur et lui faciliter l'annotation. L'usage de l'application a été mesuré de différentes manières, par des tests de performance objectifs ainsi que par des avis subjectifs d'utilisateurs, tous présentés dans cette section.

7. La Commission Nationale de l'Informatique et des Libertés a son site web accessible à l'adresse http://www.cnil.fr/english/. Sa mission est de veiller au respect des droits et libertés des individus dans l'usage des nouvelles technologies.

3.3.1 Description technique

L'application RECORDME a été créée à l'aide du kit de développement Android SDK[8] (*Standard Development Kit* en anglais) dont c'était la version n°15. La collecte des différentes sources de données est possible grâce aux classes et objets à disposition dans l'interface de programmation du SDK qui permettent la manipulation des capteurs. En particulier, il est possible de détecter ceux qui sont présents sur le smartphone, de les affecter à un enregistrement et de les configurer (par exemple, pour indiquer la fréquence d'échantillonnage). Les capteurs sont gérés par des événements : dès qu'une nouvelle valeur est disponible, elle déclenche l'appel d'une fonction. C'est cette fonction, qui permet de sauvegarder la valeur. Il est également possible de sauvegarder l'identifiant du capteur, l'horodatage de la mesure et une estimation de la précision. Les détails des données collectées par RECORDME sont décrits dans la table 3.3.

TABLE 3.3: Sources, types et fréquences des données collectées

Source	**Donnée Collectée (donnée ; type ; représentation)**	**Échantillonnage**
Accéléromètre	Accélération ; numérique ; 3 valeurs brutes sur 3 axes	Régulier, jusqu'à 100 Hz
Magnétomètre	Champ magnétique ; numérique ; 3 valeurs brutes sur 3 axes	Régulier, jusqu'à 100 Hz
Gyroscope	Vitesse de rotation ; numérique ; 3 valeurs brutes sur 3 axes	Régulier, jusqu'à 100 Hz
Luminosité	Luminosité ; numérique ; 1 valeur	Régulier, jusqu'à 100 Hz
Proximité	Proximité ; binaire nominal ; (près / loin)	Événementiel
Baromètre	Pression atmosphérique ; numérique ; 1 valeur	Régulier, jusqu'à 100 Hz
Son	ZCR, Amplitude des coefficients de DFT ; numérique ; descripteurs	Régulier, 44100 Hz
Écouteurs	Branchement ; binaire nominal ; (branché / débranché)	Événementiel
Batterie	Niveau ; binaire nominal ; (niveau faible / niveau haut) Rechargement ; binaire nominal ; (Chargeur branché / débranché)	Événementiel Événementiel
Écran	État ; binaire nominal ; (allumé / éteint)	Événementiel
Applications	Nom de l'application utilisée ; nominal ;	Événementiel
Journal d'appels	État du service ; nominal ; (exemple : en service) État de l'appel ; nominal ; (Disponible / En communication / Sonne) Numéro entrant ; numérique ;	Événementiel Événementiel Événementiel
Journal des sms	Numéro entrant ; numérique ;	Événementiel
Journal des données	État de l'échange ; nominal ; (exemple : Aucun, Entrant, Sortant) État de la connexion ; nominal ; (Données ou Wi-Fi ; Statut)	Événementiel Événementiel
Journal du Bluetooth	État ; binaire nominal ; (Allumé / Éteint) État recherche appareils ; binaire nominal : (démarré / terminé) Adresse MAC appareil détecté ; nominal ;	Événementiel Événementiel Événementiel
Journal du Wi-Fi	État ; binaire nominal ; (allumé / éteint) Identifiant borne détectée ; nominal ; Identifiant borne connectée ; nominal ;	Événementiel Régulier ou Événementiel Événementiel
Antennes relais	Identifiant antenne-relais connectée ; nominal ;	Événementiel
Localisation GPS / Wi-Fi	Coordonnées (latitude, longitude, altitude, vitesse, précision mesure) ; numérique ; une valeur pour chaque élément	Régulier ou Événementiel

L'enregistrement simultané des différentes sources est possible grâce à l'usage d'une structure propre à Android (le Service) et de la gestion simultanée des *threads*, qui sont deux moyens d'effectuer une tâche de fond. Un *thread* est créé par défaut dans le proces-

8. Le SDK est disponible en téléchargement depuis l'adresse https://developer.android.com/sdk/index.html.

sus de l'application pour gérer les actions réalisées *via* l'interface graphique. Par défaut, une nouvelle tâche est assignée au *thread* principal mais il est possible de l'affecter à un autre *thread*. Le *Service* est une classe d'objets destinée à réaliser des tâches longues en arrière-plan. Cependant, ces deux éléments ne suffisent pas pour l'enregistrement continu pendant une longue période. En effet, chaque *thread* possède un niveau de priorité de fonctionnement dont le système d'exploitation se sert pour affecter les ressources aux tâches. L'une de ses missions est de préserver les ressources. Ainsi, son comportement tend à mettre le processeur en veille au bout d'un certain temps d'inactivité de l'utilisateur, même si ces tâches fonctionnent en fond. Il est cependant possible de contourner ce problème par l'usage d'un objet dédié appelé *WakeLock* dont la fonction est de maintenir le processeur éveillé. Grâce à cela, le système ne cherche pas à passer en veille et l'enregistrement peut continuer. La contrepartie est la diminution plus rapide de la batterie.

Le transfert des fichiers sans fil est réalisé au moyen d'une connexion Wi-Fi. Le transfert démarre à la demande de l'utilisateur et lorsque la connexion à une borne a été établie. Nous avons fait appel à la bibliothèque de code Java intitulée JSch [9] qui propose un ensemble de fonctions pour la manipulation à haut niveau des objets réalisant les connexions sécurisées suivant le protocole SSH. En particulier, la bibliothèque de code propose l'envoi par protocole SCP, que nous avons choisi.

3.3.2 Interface graphique

RECORDME se compose de 3 vues principales pour l'accueil, la sélection des sources et l'annotation. La figure 3.4 illustre les deux premières, l'interface de l'annotation ayant déjà été évoquée lors de la section 3.2.2.2. L'écran d'accueil propose l'accès aux principales fonctionnalités de gestion de l'enregistrement, d'annotation et de transfert des fichiers. En outre, l'état de l'enregistrement est affiché sur cette vue. La gestion de l'enregistrement se fait au moyen des boutons de démarrage et d'arrêt sur l'écran d'accueil. Pour démarrer un enregistrement, l'utilisateur se retrouve sur la vue de la sélection des sources qui affiche les capteurs disponibles sur le smartphone. En cours d'enregistrement, les capteurs activés sont marqués par un segment vert sous leur nom (alors que le segment est gris avant un enregistrement ou lorsqu'ils ne sont pas sélectionnés). Ainsi, le participant connaît l'état précis des capteurs en cours d'enregistrement. L'arrêt d'un enregistrement et le transfert de fichiers se lancent depuis l'écran d'accueil et n'ouvrent pas de nouvelle vue.

En plus des vues présentées, l'application dispose de notifications qui s'affichent dans la barre dédiée, en haut de l'écran, pour informer l'utilisateur des actions principales. Quand un enregistrement est actif, la notification de la figure 3.5.a est affichée dans la barre et permet de retourner directement à l'application en cliquant dessus. Également, lors du transfert des fichiers *via* la liaison sans fil, le statut est affiché avec les icônes de la figure 3.5.b.c.d. qui indiquent trois états possibles du transfert : en cours, succès du transfert d'un fichier et échec du transfert d'un fichier.

9. JSch (*Java Secure Channel*) est une bibliothèque de code en Java pour l'établissement de connexions sécurisées suivant le protocole SSH (*Secure Shell*).

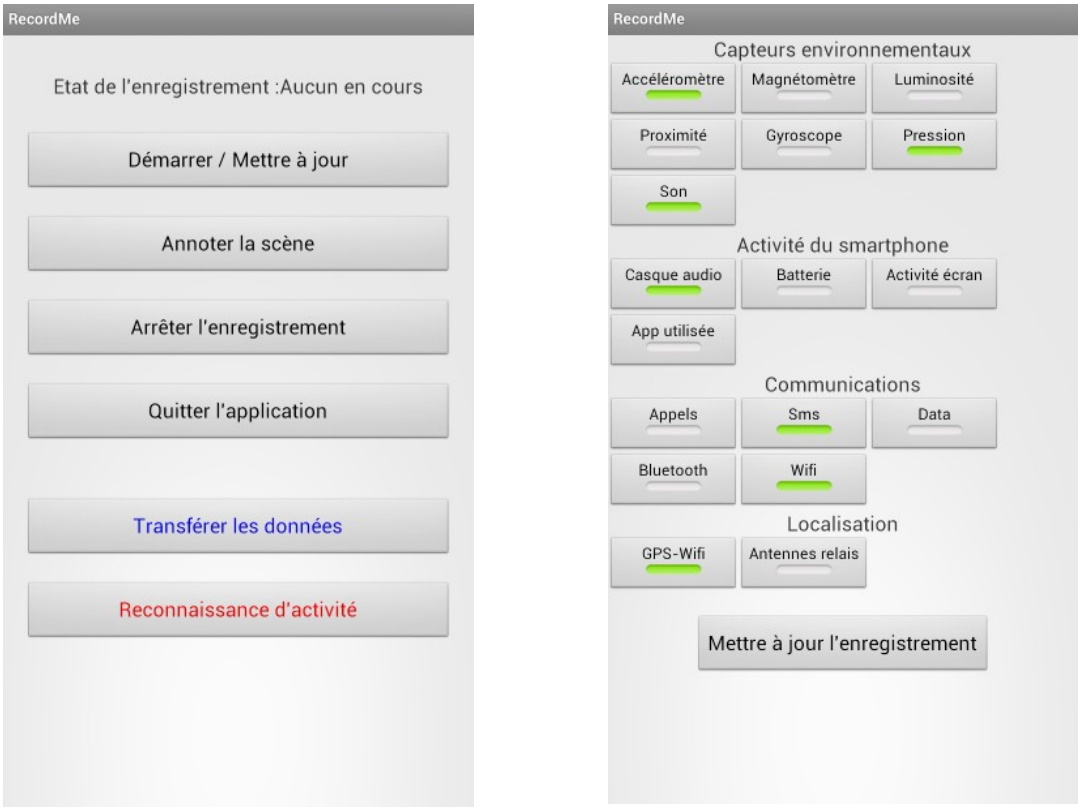


FIGURE 3.4: Captures d'écran de RECORDME avec de gauche à droite : a) l'écran d'accueil et b) l'écran de sélection des sources.

FIGURE 3.5: Icônes des notifications de l'application RECORDME avec à gauche a) l'icône d'enregistrement en cours ; puis les icônes de l'état du transfert des données *via* la liaison sans fil, respectivement b) en cours, c) réussi et d) échoué.

3.3.3 Informations pratiques et mesures

Les paragraphes suivants offrent un regard critique sur l'application par le biais de tests de performance, d'avis d'utilisateurs et de dysfonctionnements remarqués.

3.3.3.1 Tests et performances

L'application a été testée avec succès sur plus d'une dizaine de téléphones de différentes marques (Acer, Google, Motorola, Samsung, Sony, Wiko, MTT [10]), de différentes gammes et sur deux versions d'Android (2.3 et 4.0). Des tests ont été menés pour estimer l'impact sur le smartphone du fonctionnement de l'application, les résultats sont présentés dans la table 3.4. Les tests ont été effectués sur 1 smartphone de haute gamme (le Samsung Galaxy

10. Acronyme de Mobile Tout Terrain. La marque se concentre sur des appareils résistants à l'immersion, à l'infiltration de poussière et de neige et vise ainsi des conditions d'utilisation extérieures. Pour plus de détails, voir le site https://www.mobiletoutterrain.com/FR/

SIII).

Sources	Charge proc.	Utilisation batterie	Stockage (Mo)
Accéléromètre (A)	8 %	-5 %	3,9
Microphone (M)	29 %	-5 %	31,8
Wi-Fi (W)	< 1 %	-5 %	< 1
GPS (G)	< 1 %	-16 %	< 1
(A + M + W + G)	30 %	-16 %	36
Tous	**32 %**	**-16 %**	**54**

TABLE 3.4: Performances mesurées pour l'enregistrement de différentes sources *via* RECORDME

La charge du processeur a été mesurée au moyen de la commande *top* sur le téléphone et accessible par le programme de débuggage *adb*, livré avec le kit de développement. Celui-ci permet entre autres d'ouvrir un terminal sur l'appareil et d'y exécuter des commandes. La commande est exécutée 10 fois pendant 10 minutes lors de l'enregistrement de sources. Chaque réponse de la commande représente la charge du processeur moyennée sur la période d'une minute. Nous avons moyenné les dix réponses fournies par la commande. Comme on peut le voir sur la table, l'enregistrement représente presque le tiers de la charge totale du processeur. Cela peut s'expliquer par la fréquence d'échantillonnage élevée et le calcul des descripteurs (en particulier, les calculs de transformée de Fourier nécessitent beaucoup d'opérations).

La mesure de la consommation de la batterie est plus délicate car il n'y a pas de méthode pour calculer précisément la consommation d'un processus. Cependant, nous pouvons réduire l'impact de certaines sources connues pour leur forte consommation d'énergie. Ainsi, les différentes radios (antennes-relais, Bluetooth, Wi-Fi, GPS) doivent être coupées si elles ne servent pas à l'expérimentation. Également, l'écran doit être allumé le moins possible pendant la mesure pour limiter la consommation d'énergie. À l'inverse, l'obligation dans l'application RECORDME de maintenir le processeur éveillé est une source de consommation importante, imputable à l'application. Ainsi, les mesures ont été effectuées suivant ces considérations et ont consisté à calculer la différence du pourcentage de charge restant avant et après un enregistrement d'une heure.

L'espace de stockage nécessaire à l'enregistrement des différentes ressources est obtenu par calcul et a été vérifié sur les téléphones. Les chiffres indiqués représentent l'espace occupé par un fichier contenant une heure d'enregistrement. Les données sonores occupent le plus d'espace, tout en restant raisonnable avec à peine plus de 30 Mo d'espace alloués. L'enregistrement simultané de toutes les sources requiert un peu plus de 50 Mo pour une heure d'enregistrement. Cette taille est raisonnable et permet d'envisager des enregistrements d'une dizaine d'heures avec la plupart des cartes externes de stockage vendues sur le marché (dont la taille se chiffre très souvent en une dizaine de giga-octets). L'usage de la seule mémoire interne du téléphone pour le stockage des fichiers est également possible, mais il faut s'assurer d'avoir au minimum 500 Mo libres pour pouvoir envisager dix heures d'enregistrement.

3.3.3.2 Avis d'utilisateurs

Un entretien a été conduit avec deux participants à la collecte. Ces participants sont membres de l'équipe mais n'étaient pas impliqués dans le projet, leur retour nous semble donc important puisqu'ils peuvent être assimilés à des participants extérieurs. L'entretien s'est déroulé sous forme d'une conversation, orientée par des questions prédéfinies. Parmi elles, les points suivants ont été abordés :

— Impression générale de l'application (fonctionnement, interface) ;

— Comportement et annotations (compréhension des annotations, changement des habitudes, oubli d'annotations) ;

— Sentiment d'intrusion dans la vie privée (sentiment de surveillance, modification des conditions d'enregistrement en conséquence) ;

— Comportement du smartphone (ralentissements, décharge plus rapide de la batterie, manque d'espace de stockage) ;

— Dysfonctionnements notables (description, conséquences telles que fonction non-réalisée, gêne, terminaison brutale d'une application).

L'impression générale des participants interrogés est bonne. L'interface leur a paru simple à utiliser et fonctionnelle, le comportement de l'application a été stable à l'exception d'un problème rencontré par l'un des participants (impossibilité de transférer des fichiers *via* l'interface de transfert des données sans fil). Le procédé d'annotation ainsi que les contextes à annoter ont été globalement bien compris par les participants. Par ailleurs, les participants n'ont pas eu l'impression d'être surveillés par l'application pendant les enregistrements.

3.3.3.3 Dysfonctionnements remarqués

Au cours des nombreux tests et enregistrements réalisés, nous avons relevé deux comportements problématiques. Le premier est celui rapporté par les participants et concerne l'envoi de fichiers *via* la fonctionnalité de transfert sans fil. Après investigation, nous avons remarqué que l'application a tenté d'établir une connexion au serveur mais celle-ci a été coupée. Une discussion avec les administrateurs du serveur suggère qu'il pourrait s'agir d'un blocage du fournisseur d'accès à internet pour empêcher l'utilisation du protocole SCP sur certains ports de communication.

Le second problème concerne les données sonores enregistrées. Nous avons remarqué des discontinuités dans la séquence d'échantillons. Nous expliquons ce problème par la charge de l'enregistrement sur le processeur, confirmée par les mesures rapportées précédemment. En imaginant un fonctionnement normal avec les différentes radios allumées et le fonctionnement d'applications tierces, il est raisonnable d'imaginer une charge importante du processeur, pouvant mener à l'interruption momentanée d'applications, ce qui peut conduire à la perte de tableaux d'échantillons. Cependant, nous avons pu détecter la plupart des discontinuités grâce à l'horodatage des échantillons.

3.3.3.4 Exemple de signaux collectés

Nous donnons à titre indicatif un exemple de plusieurs signaux collectés par un volontaire au cours d'un enregistrement de plusieurs heures afin d'illustrer les possibilités d'analyse de ces signaux. De haut en bas, les signaux correspondent à l'amplitude de l'accélération exprimée en $m.s^{-2}$; l'usage courant des applications, dont les noms sont remplacés par des indices pour les rendre anonymes ; l'état de l'écran du smartphone ; l'évolution des antennes relais visitées au cours du temps, elles aussi anonymes ; et la séquence des environnements visités.

On peut alors imaginer des analyses par comparaison ou corrélation des motifs des différents signaux. Des corrélations peuvent être observées entre les données et les scènes (par exemple, l'usage d'une application qui se produit souvent dans une scène particulière). Cette information peut ensuite être exploitée dans un système de reconnaissance. Aussi, les données de fonctionnement, telles que l'allumage de l'écran de l'appareil, peuvent être exploitées comme des références ou des annotations complémentaires. Par exemple, si l'on souhaite analyser les accélérations dans les scènes afin d'étudier les mouvements effectués, on peut considérer les périodes où l'écran est éteint, suivant l'hypothèse que durant ces périodes, l'appareil n'est pas utilisé. Les périodes avec l'écran éteint sont identifiables grâce aux données collectées.

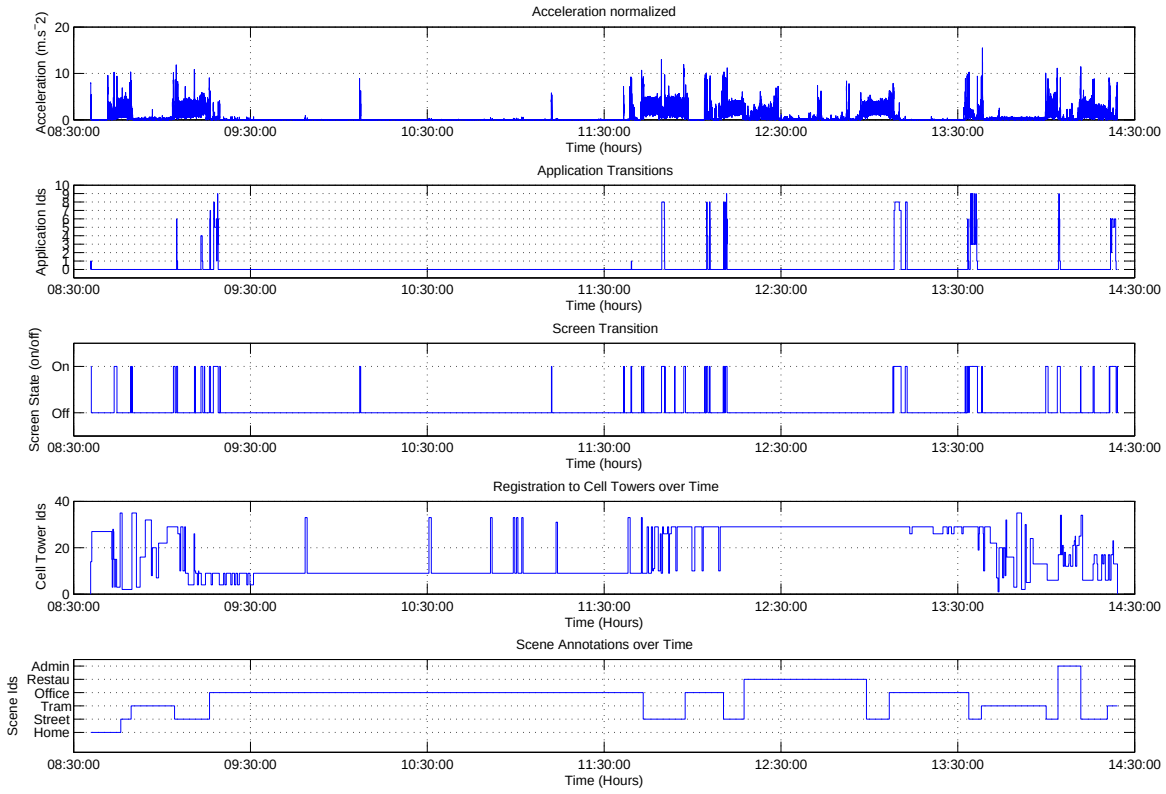


FIGURE 3.6: Exemples de signaux collectés

3.4 Les collectes

La section présente les deux collectes réalisées avec l'application RECORDME. La collecte de scènes est la plus importante des deux. Elle suit le protocole présenté précédemment pour enregistrer et annoter des données réelles. La seconde collecte est dédiée aux activités physiques simples, postures et différentes positions du smartphone, qui n'ont pu être annotées au cours de la première collecte.

3.4.1 La collecte de scènes

C'est la collecte principale effectuée avec l'application RECORDME. Six volontaires y ont participé, sur une période de plusieurs mois au cours du printemps et de l'été 2014. La plupart des enregistrements ont été effectués dans la région de Grenoble ; quelques uns proviennent d'autres régions de France et de l'étranger (notamment de Budapest, Dublin et Singapour). La durée totale cumulée des enregistrements est de 575 heures, réparties en 82 enregistrements. Parmi ces heures de données, certaines sont marquées de l'étiquette d'incertitude ; le total d'heures dont les annotations ont été validées se monte à 536 heures.

Dans la table 3.7, nous rapportons les durées cumulées et nombre d'instances suivant les scènes visitées et les sources enregistrées. Pour les sources continues, la quantité est exprimée en heures tandis que pour les sources événementielles, elles est exprimée en nombre d'événements enregistrés.

On remarque que l'accéléromètre et le magnétomètre sont les capteurs les plus présents. La présence moindre du gyroscope et du baromètre s'explique par l'absence de ces capteurs de certains téléphones (en particulier, ceux d'entrée de gamme). Le son est également moins présent. Cela s'explique par les mesures de protection mises en place (telles que l'arrêt de l'enregistrement du son en cas d'appel). Également, il est possible que certains utilisateurs aient préféré couper l'enregistrement afin de se sentir plus tranquille.

Le nombre d'instances des sources événementielles est très hétérogène allant de quelques unités ou centaines pour le Bluetooth ou l'indicateur de batterie, à plus de 20000 pour le journal des données. Les coordonnées géographiques du GPS et les identifiants des bornes Wi-Fi peuvent être considérées comme des données régulières. Si l'on multiplie par la fréquence théorique d'acquisition des coordonnées qui est d'1 Hz, on obtient une "durée" de 519446 secondes, soit approximativement 144 heures d'acquisition.

Les différentes informations fournies par une source événementielle (comme par exemple, l'état du Wi-Fi et les identifiants des bornes détectées et finalement connectées) ne sont pas distinguées dans les valeurs de la table. Ainsi, ces valeurs représentent le nombre total d'événements mesurés pour une source événementielle.

3.4.2 La collecte d'activités physiques et de positions du smartphone

Cette collecte complète la collecte de scènes précédente par l'enregistrement de données et l'annotation d'activités physiques simples et de postures, suivant différentes posi-

Scène	Durée (h)	Durée (%)
Métro	0,4	0,1
Moto	0,8	0,1
Avion	2,1	0,4
Train	5,5	0,9
Tram	8,0	1,4
Bus	9,6	1,7
Commerce	14,0	2,4
Montagne	14,0	2,4
Bateau	14,8	2,6
Voiture	20,1	3,5
Bureau, pause	23,7	4,1
Restaurant, bar	29,4	5,1
Rue	37,1	6,4
Incertain	39,2	6,8
Bureau, réunion	54,3	9,4
Domicile	81,3	14,1
Bureau, travail	221,4	38,5
Total	**575,6**	**100,0**

Sources Régulières (heures)	
Source	**Environnements**
Accéléromètre	391.9
Magnétomètre	391.9
Gyroscope	160.4
Luminosité	337.9
Proximité	
Baromètre	127.3
Son	253.1
Sources Événementielles (unités)	
Écouteurs	168
Batterie	172
Écran	5704
Applications	3636
Journal d'appels	5005
Journal de sms	455
Journal des données	21554
Journal du Bluetooth	4
Journal du Wi-Fi	109556
Antennes relais	16914
Localisation GPS/Wi-Fi	519446

FIGURE 3.7: Représentation (à gauche) des durées cumulées des scènes enregistrées ; et (à droite) des durées et instances des sources enregistrées

tions du smartphone (Blachon et coll. (2014a)). La collecte diffère également par le contrôle plus important pendant les sessions d'enregistrement. Le corpus obtenu est exploité dans une expérimentation décrite dans la section 6.2.1 du chapitre 6.

3.4.2.1 Description générale de la collecte

La collecte vise l'enregistrement des éléments suivants :

— activité physique simple : la marche, la montée et descente d'escaliers, la course et le saut ;

— usages du téléphone : pas d'usage (téléphone posé), appeler, utiliser une application, écrire et envoyer un sms et écouter de la musique ;

— postures immobiles : assis, debout et allongé ;

— position du smartphone : tenu à la main, gardé dans la poche et gardé dans le sac.

Pour cela, des scénarios ont été mis en place, supervisés par un expérimentateur au cours de l'enregistrement qui guide le volontaire dans la séquence des actions à effectuer et reporte les annotations. Le volontaire porte simultanément quatre téléphones, placés dans les différentes positions (deux téléphones se situent dans le sac).

Un scénario se compose d'une séquence de postures et activités physiques à effectuer. Pour chacune d'elles, la position des différents smartphones est spécifiée. L'usage du téléphone tenu dans la main est également mentionné. Au-delà de ces spécifications, le volon-

taire est libre dans la réalisation des activités et postures et dans le temps passé dans chacune d'elles.

Un extrait d'un scénario est rapporté dans la table 3.5. Chaque téléphone est associé à un numéro. La dernière colonne de la table indique la position des trois premiers téléphones. Le quatrième téléphone est resté au fond du sac.

Activité ou posture	Usage	Positions smartphones
Assis	Posé	n°1 main ; n°2 poche ; n°3 sac
Assis	Appel	n°1 main ; n°2 poche ; n°3 sac
Assis	Non	n°1 main ; n°2 poche ; n°3 sac
Debout	Non	n°1 main ; n°2 poche ; n°3 sac
Marche	Non	n°1 poche ; n°2 sac ; n°3 main
Escaliers	Non	n°1 poche ; n°2 sac ; n°3 main

TABLE 3.5: Extrait d'un scénario joué par l'un des volontaires

La synchronisation des annotations avec les indications temporelles des données, elles-mêmes issues des différents smartphones, est possible grâce l'alignement d'un événement extérieur capturé par les appareils. Il s'agit d'aligner les appareils sur une table après avoir démarré l'application de collecte, puis d'asséner un coup sur la table au moyen de la main ou d'un objet laissé tombé. La force engendrée crée une accélération suffisamment forte et ponctuelle pour être détectée. Cet événement crée aussi un bruit fort et court, identifiable pour synchroniser les données sonores. Le smartphone de l'expérimentateur employé pour les annotations est également synchronisé de cette manière.

La synchronisation des données consiste à rechercher le pic d'énergie dans les signaux d'accélération et de son et à considérer les indicateurs temporels associés comme origine du temps.

Le reste du protocole est similaire à celui qui a été présenté dans la section 3.2.5 : le volontaire est accueilli et informé des données collectées et traitements effectués ; le post-traitement des données indexe les enregistrements dans la base et recherche les dysfonctionnements éventuels pour détecter des discontinuités dans les signaux.

3.4.2.2 Les données

La collecte a mobilisé 19 volontaires, qui ont chacun effectué une session d'enregistrement. La durée moyenne d'une session est de 13 minutes. Les quatre smartphones étaient portés simultanément par les volontaires. La durée cumulée des enregistrements est de 988 minutes (soit environ 16 heures). Les sources enregistrées se limitent au journal du fonctionnement du téléphone, à l'accéléromètre et au microphone.

Les quatre téléphones employés sont un Motorola Defy Mini, un MTT, un Samsung Galaxy S2 and un Google Nexus 4. Les deux premiers représentent l'entrée de gamme tandis que les deux derniers sont de meilleure qualité. Nous avons constaté de nombreuses discontinuités dans les données issues du Motorola et du MTT, qui nous ont conduits à les retirer complètement du corpus de données. Nous avons également remarqué un dysfonctionnement

dans l'enregistrement des trois premiers volontaires, et avons retiré les données de leurs enregistrements en conséquence. Finalement, le corpus exploitable se résume à 16 volontaires enregistrés par deux téléphones, et totalise 408 minutes cumulées de données (approximativement 7 heures). La table 3.6 présente les durées cumulées des postures et activités. Nous rapportons également l'étape du téléphone posé sur la table, qui représente les passages pendant lesquels le téléphone n'était pas porté par l'utilisateur.

	Étiquette	**Durée (min)**	**Durée (%)**
Act. phys.	Sauter	4,9	1,2
	Courir	14,0	3,4
	Escaliers	24,4	6,0
	Marcher	93,8	23,0
Postures	Allongé	36,0	8,8
	Debout	69,9	17,1
	Assis	141,0	34,6
	Téléphone posé	24,1	5,9
	Total	**408,1**	**100,0**

TABLE 3.6: Distribution des durées des activités enregistrées

3.5 Bilan de la collecte

Le chapitre rapporte une des premières étapes de la thèse pour la mise en place d'une collecte de données. Nous avons d'abord défini les caractéristiques des données en déterminant les scènes d'intérêt souhaitées, les sources de données intéressantes et les caractéristiques précises des données : réelles et annotées ; provenant de différentes sources et synchronisées ; et continues. Les conditions énoncées sont strictes et, à notre connaissance, aucun corpus disponible ne satisfait l'ensemble des caractéristiques. La même observation est dressée pour les outils de collecte. Cela a justifié la réalisation d'un outil et d'une collecte.

La réflexion sur le protocole pour la collecte s'est d'abord confrontée à la difficulté d'effectuer une acquisition de données réelles *in vivo* et de collecter des annotations de qualité. La solution que nous proposons repose sur l'auto-annotation et l'évaluation *a posteriori* des annotations obtenues. L'auto-annotation doit être exécutée dans un cadre favorable, c'est pourquoi notre méthode consiste à informer les volontaires des concepts à annoter ainsi que des détails de la collecte. L'outil a aussi été pensé pour faciliter la tâche des volontaires. L'évaluation repose sur des hypothèses formulées sur les concepts à annoter et sur l'exploitation de données objectives de géo-localisation. Le second problème pour l'établissement du protocole a concerné la sécurité des données et le risque d'intrusion dans la vie privée. Après l'identification de ces risques tels que ceux portant sur les contenus sensibles des données ou le risque d'interception, nous avons proposé des solutions pour les phases d'acquisition, de transfert et de stockage de longue durée.

La réalisation de la collecte a aussi nécessité la conception de l'outil d'acquisition appelé RECORDME. Les caractéristiques des données et du protocole ont influencé son développement. L'architecture permet, entre autres, un enregistrement continu et simultané de plusieurs sources et la collecte d'indicateurs temporels des données pour une synchronisation ultérieure. L'interface d'annotation a été pensée pour faciliter la tâche aux volontaires.

Afin de compléter la description de l'application, nous avons également présenté des résultats d'évaluations de l'application. Elle a été testée sur une dizaine de smartphones de marques et de gammes différentes ainsi que sur deux versions différentes du système d'exploitation Android. Les tests de performance indiquent une charge importante du processeur, principalement due à l'enregistrement et au traitement du son. Cette observation pourrait expliquer les dysfonctionnements ponctuels de l'application et les discontinuités observées dans certains enregistrements sonores. Des entretiens avec les utilisateurs ont été menés. Ils indiquent une satisfaction du comportement de l'application, une facilité d'utilisation, très peu de gêne ou de ralentissement dans le fonctionnement normal du téléphone et pas de sentiment d'observation.

Deux collectes ont été effectuées avec l'application. La première porte sur l'acquisition de données et d'annotations dans les scènes d'intérêt. Elle totalise 575 heures de données (dont les annotations de 536 heures ont été validées), enregistrées par 6 volontaires principalement dans la région de Grenoble mais également à l'étranger (à Budapest, Dublin et Singapour). La seconde collecte complète la première par l'acquisition d'activités physiques simples, de postures et de positions du smartphone, qu'il était contraignant d'acquérir dans la première collecte. La seconde collecte est supervisée, composée de scénarios suivis par les volontaires pour effectuer les différentes actions. Un expérimentateur se charge des annotations. Le corpus total cumulé exploitable contient environ 7 heures de données et se compose de 16 sessions d'enregistrements, chacune effectuée par un volontaire et enregistrée pour différentes positions du smartphone. Ces corpus sont maintenant à disposition pour différents traitements.

Pour conclure, la démarche mise en place et les corpus collectés sont satisfaisants et répondent aux problèmes de diversité des sources, de réalisme des données, d'anonymat des volontaires et de réalisation et de validation de l'annotation.

CHAPITRE 4

Le modèle de scène

Ce chapitre est le cœur du manuscrit. D'abord, nous abordons le problème de la définition de scène en exploitant les résultats des deux chapitres précédents. La constitution d'un corpus de données a permis la collecte d'annotations de scènes qui sont étudiées en vue de comprendre la perception des scènes par les participants. L'état de l'art a fourni des éléments sur la représentation du contexte dans l'intelligence ambiante et des exemples de travaux dans lesquels les scènes sont représentées.

La combinaison de ces résultats est présentée dans ce chapitre pour en retirer une définition de scène. Nous décrivons également comment la définition s'applique dans les expérimentations que nous avons réalisées. Pour cela, nous formulons plusieurs hypothèses sur les expérimentations, qui tiennent également compte des contraintes de collecte et des objectifs industriels.

À partir de ces éléments, nous pouvons introduire et justifier les travaux réalisés au cours de la thèse en vue de répondre aux problèmes posés. Le premier problème est la reconnaissance de scène, pour lequel nous proposons plusieurs solutions de référence. Le second problème porte sur l'imprécision de la définition de scène. Alors que la première partie du chapitre aborde le problème à partir des annotations et des travaux de l'état de l'art, nous décrivons une autre approche basée sur l'interprétation des données et des concepts qui en émergent après regroupement non-supervisé.

4.1 Définition d'une scène

L'exemple de scène fourni par le partenaire industriel évoque une situation de réunion de travail. On extrapole de cet exemple que les scènes d'intérêt semblent être des moments particuliers de la vie quotidienne des utilisateurs. Également, cet exemple éclaire sur le niveau de granularité de la scène : la réunion entière représente la scène ; on ne s'intéresse pas à décrire chaque sous-événement dans la réunion (écoute, prise de parole, etc).

Cet exemple est bien entendu trop restreint pour apporter un éclairage sur ce qu'est une scène. L'état de l'art des domaines tel que la reconnaissance de scènes sur smartphone n'apporte également pas une réponse claire. À partir des annotations produites par les volontaires, nous souhaitons déterminer les scènes d'intérêt, notamment par leur fréquence d'apparition ou leur durée. Nous espérons également mieux définir la perception des scènes, grâce aux descriptions fournies dans les annotations libres. La confrontation des résultats

TABLE 4.1: Annotations des scènes issues des listes prédéfinies

	Scène	Inst.		Scène	Inst.
Extérieur	Rue - marche	224	Bureau	Bureau - travail ordi.	83
	Parc	5		Bureau - réunion	37
	Plage	4		Bureau - pause	19
	Mer et port	3		Bureau - déjeuner	4
Transports	Tramway	44	Dom.	Domicile - disponible	58
	Bus	28		Domicile - indisponible	11
	Voiture - conducteur	9		Maison d'un proche	11
	Voiture - passager	19	Bâtiments publics	Commerce - courses	21
	Moto	8		Administration	4
	Métro	8		Restaurant, bar	22
	Avion	4		Musée	8
	Train	6		Théâtre	4
	Vélo	27		Gare	2
	Bateau	2		Aéroport	3

de ces deux approches à des définitions issues de différents domaines permet d'aboutir à des contraintes identifiables pour la définition d'une scène.

4.1.1 Analyse des annotations humaines des scènes

L'étude des annotations est effectuée d'abord en considérant les annotations obtenues à partir de listes prédéfinies dans l'interface d'annotation de l'application de collecte. Nous confrontons les premières observations émises aux annotations libres fournies par les participants. Nous complétons l'analyse par l'étude du rapport de durée et fréquence des annotations.

Étude des annotations guidées

L'ensemble des annotations produites par les volontaires s'élève à 756 éléments. Parmi elles, 678 proviennent des listes prédéfinies et 78 des annotations libres. La table 4.1 résume les annotations issues des listes ainsi que leur nombre d'instances. Une instance de scène correspond à une période continue pendant laquelle l'utilisateur se trouve dans celle-ci. Elle se termine lorsqu'il change de scène (et ainsi commence l'instance de la scène suivante).

La table met en évidence des scènes fréquentes, identifiées par le nombre d'instances élevé. Parmi elles, on trouve la marche dans la rue ; les transports tels que le tramway, le bus, la voiture ou le vélo ; le bureau et la distinction de différentes activités effectuées ; le domicile ; et différents bâtiments dits publics avec, en tête, les commerces et les restaurants ou bars. La forte représentation de ces scènes s'explique, d'une part, par la perception commune de celles-ci par plusieurs volontaires et, d'autre part, par la fréquence de leurs réalisations pour un volontaire, ce qu'on imagine aisément dans un environnement urbain.

L'observation de la table montre également différents niveaux de granularité dans le regroupement des scènes. Le premier niveau observable est indiqué par les séparations propo-

sées dans la table. Les groupes qui en découlent n'ont pas émergé naturellement des annotations, mais proviennent de la catégorisation des étiquettes dans l'interface d'annotation. Cependant, ces groupes sont justifiés par les caractéristiques communes des scènes qu'ils contiennent et leur "poids" dans les annotations, représenté par la présence de scènes fréquentes dans leur composition. Certains de ces groupes peuvent être fusionnés. Ainsi, les groupes du bureau, du domicile et des bâtiments publics contiennent tous des scènes en intérieur, par opposition au groupe de scènes en extérieur. Les transports représentent un cas particulier car ils ne sont pas des scènes fixes. Mais on peut les considérer comme des lieux de vie par la durée qu'on y passe et les activités effectuées. Un niveau d'interprétation intermédiaire entre ces groupes et les scènes peut également émerger. Par exemple, parmi les transports, on peut distinguer ceux qui sont motorisés de ceux qui ne le sont pas. Le vélo est le seul représentant de ce dernier cas. De manière identique, les scènes du bureau ou des bâtiments publics peuvent être différenciées suivant l'activité exercée.

Pour terminer l'étude de la table 4.1, on peut remarquer que la plupart des scènes sont décrites par une étiquette symbolique qui représente un lieu. On remarque également que le lieu n'est parfois pas suffisant pour décrire la scène et, dans ce cas, l'usage d'une activité ou d'une attitude vient compléter la description (par exemple pour le lieu du bureau). Nous avons conscience que ces observations sont émises sur l'étude d'annotations guidées par des listes prédéfinies, c'est pourquoi nous étudions les annotations libres dans la suite.

Étude des annotations libres

Les annotations libres collectées sont résumées dans la table 4.2. Nous les avons regroupées dans différentes catégories pour faciliter leur étude. Ainsi, on remarque les trois catégories de la moitié gauche de la table qui confirment l'usage limité des étiquettes symboliques de lieu ou d'action. La première section de la table regroupe les annotations qui décrivent les attitudes ou actions dans un lieu. Par exemple, on y trouve des périodes de pause ou d'attente, dans les lieux de la rue ou du bateau qui distinguent ces scènes de la marche dans la rue ou de la posture assise dans le bateau. Également, il y a le cas de l'hôtel et de l'aéroport où l'on peut pratiquer des activités, telles que dormir ou prendre un café, qui sont associées à d'autres lieux.

La seconde section regroupe des descriptions de lieux. Ainsi, la scène du restaurant peut varier suivant la cuisine proposée et l'ambiance créée. Des occasions particulières comme la retransmission d'une rencontre sportive peuvent aussi changer la scène et sa perception. Une étiquette générique de lieu peut être complétée pour mieux décrire la scène. Dans la table, le lieu de "commerce" est complété par la notion de "halles" qui laisse imaginer un marché couvert avec un bruit élevé et une ambiance particulière.

La troisième section regroupe des descriptions de l'activité de réunion. On remarque des précisions apportées sur le lieu (tel que l'auditorium ou la salle de classe) ou sur le type de réunion (présentation, conférence, ou discussion). Ces précisions changent la perception de la réunion et confirment que l'usage seul d'une étiquette symbolique pour décrire l'activité n'est parfois pas suffisant.

TABLE 4.2: Résumé des annotations libres obtenues dans la collecte de scènes

Annotations	Nombre
Description d'attitude ou d'action dans un lieu	*20*
Bateau, dont passager, assis, dormir et lever	4
Rue, dont	5
Assis sur un banc	1
Pause	2
Faire la queue	2
Aéroport, café	1
Hôtel, dont	7
Hôtel, Disponible	3
Hôtel, chambre, disponible	1
Hôtel, chambre, travail	3
Domicile, dont	3
Domicile, marche	2
Domicile, téléphone en charge	1
Description du lieu	*13*
Restaurant, dont	6
Pizzéria	2
Match	2
Terrasse	2
Commerce, halles	2
Extérieur	5
Extérieur de #lieu	3
Sur le port	1
Sur le quai	1
Description de l'activité	*9*
Réunions, dont	9
Classe, donner un cours	1
Conférence	2
Exposé	2
Auditorium, réunion	1
Auditorium, présentation	1
Réunion, #lieu	2

Annotations	Nombre
Séquences	*12*
Avant aéroport puis bus	1
Arrivé à destination du lieu de vacances	2
Arrêt commerce puis vélo	1
Marche dans la rue puis assis puis marche	3
Marche dans la rue puis musée puis terrasse musée	1
Marche en extérieur puis musée	1
Marche en extérieur puis métro	1
Marche dans la rue puis hôtel	1
Marche dans la rue puis pub	1
Environnements imprévus	*4*
Aire de sport, faire du sport	1
Office du tourisme	1
Garage souterrain	1
Salle de fête, anniversaire	1
Macro-environnements	*20*
#Lieu de travail	1
Véhicule	6
Voiture	9
Intérieur	2
Dehors	2

Par la suite, les trois sections de la moitié droite de la table éclairent d'autres aspects de la perception des scènes. D'abord, la section intitulée "Séquences" présente des séquences d'annotations. Celles-ci sont la solution choisie par des volontaires après l'oubli d'une ou plusieurs annotations de scènes. L'intérêt de leur étude réside dans la forme concise de la description de la séquence, qui met en avant les scènes pertinentes pour le participant qui l'a écrite. La plupart des séquences décrivent des déplacements puis des lieux atteints tels que le métro, le musée ou l'hôtel. On remarque l'association du métro (le transport) à la station (le lieu) sous la même étiquette. Ainsi, certaines transitions entre scènes ne sont pas retenues ou sont associées à la scène précédente ou suivante.

La seconde section de la moitié droite de la table indique des environnements imprévus dans les listes prédéfinies et montre la grande variabilité des scènes suivant les utilisateurs. Cette variabilité de scènes représente une difficulté pour la création d'un système de reconnaissance de scènes qui doit, d'une part, s'adapter au "profil de scènes" d'un utilisateur et, d'autre part, intégrer de nouvelles scènes "découvertes" au fil des situations rencontrées.

Enfin, la dernière section présente des étiquettes à des niveaux de granularité plus élevés que ceux prévus dans les listes. Les transports sont parfois décrits simplement par des véhicules ; et les scènes par leur environnement en intérieur ou en extérieur. Ces observations

suggèrent la perception variable des scènes suivant les personnes ou les conditions. L'usage de plusieurs niveaux d'abstraction pour la description d'une scène pourrait permettre l'unification ou la généralisation de scènes inconnues ou aux perceptions différentes.

Étude des durées des instances de scènes

La figure 4.1 illustre le rapport du nombre d'instances et des durées cumulées pour des scènes fréquentes repérées dans la table 4.1. L'étude de la figure montre des rapports très différents de durée et de nombre d'instances des scènes, que nous exprimons par la durée moyenne d'une instance. Par exemple, le travail au bureau est une scène fréquente et longue, si l'on en juge par durée moyenne d'une instance qui est de 2 h 40 min environ. À l'inverse, les durées moyennes d'instances des scènes du tramway, du bus et de la rue sont relativement courtes en comparaison avec respectivement 11 minutes, 20 minutes et 10 minutes. Le graphique de la figure 4.1 montre que les scènes peuvent varier en durée entre elles. Cela suggère que, pour répondre au problème de la reconnaissance de scène, la considération du temps ou de la durée peut être pertinente. L'observation de scènes courtes est pertinente pour la stratégie de reconnaissance à adopter. Cela suggère que le temps ou la durée d'une scène pourraient être pris en compte dans le système de reconnaissance.

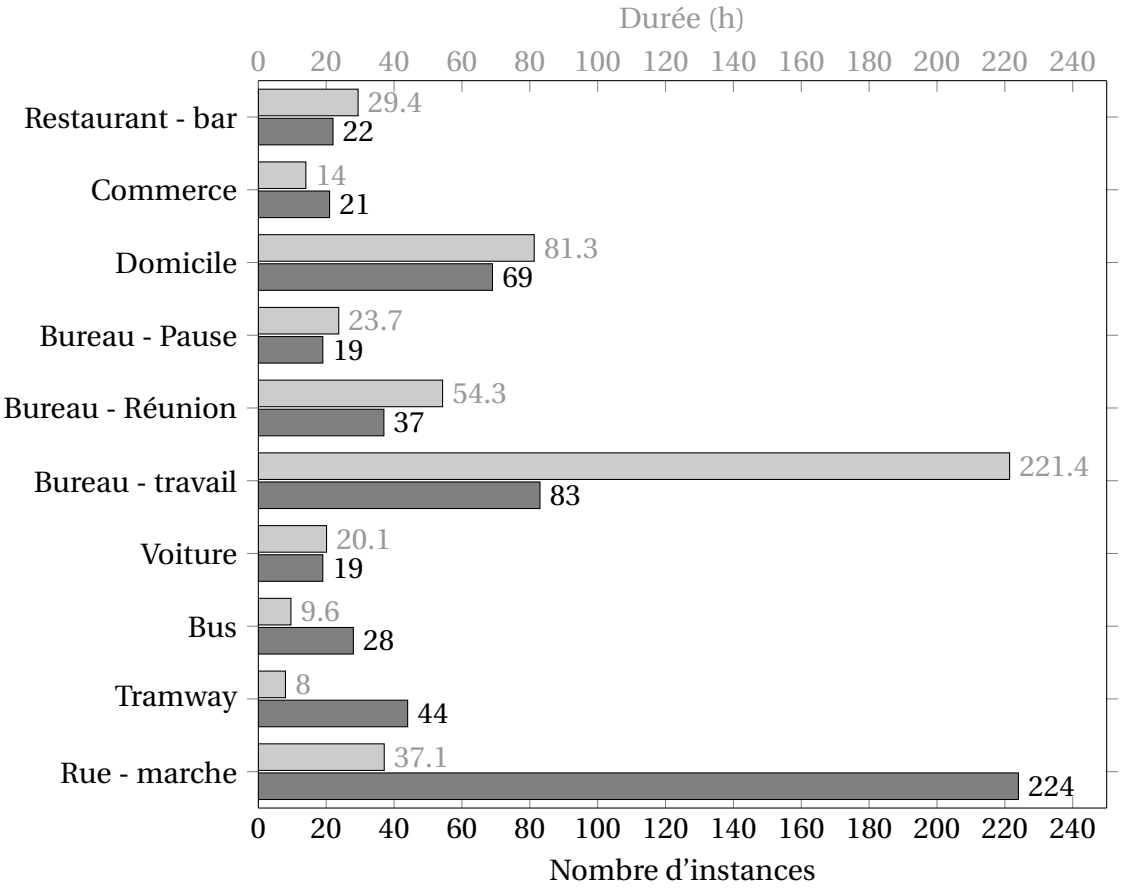


FIGURE 4.1: Nombre d'instances et durées cumulées des classes de scènes collectées

Bilan de l'analyse des annotations collectées

Nous terminons l'analyse des annotations humaines par la présentation d'un arbre de catégorisation des scènes dans la figure 4.2 construit à partir des annotations collectées et des observations faites dans cette section. La structure des branches de l'arbre est composée de lieux aux étiquettes symboliques dont on a vu qu'ils sont la première information de description de la scène. Les niveaux intermédiaires dans l'arbre indiquent les différents niveaux d'interprétation. Les niveaux choisis proviennent des différents regroupements possibles de la table 4.1. Les feuilles contiennent à la fois les scènes fréquentes repérées et d'autres scènes, moins fréquentes, montrant ainsi la grande variabilité des scènes. Cela indique également qu'une distinction peut encore être opérée sur les feuilles, mais qu'elle nécessite plus d'informations. Par exemple une caractérisation du lieu ou une description de l'action sont des éléments mis en avant dans l'étude des annotations.

Cette représentation sous forme d'arbre est intéressante pour la définition d'une scène, mais se montre limitée pour l'implémentation, en particulier dans le cas de l'adaptation à d'autres scènes.

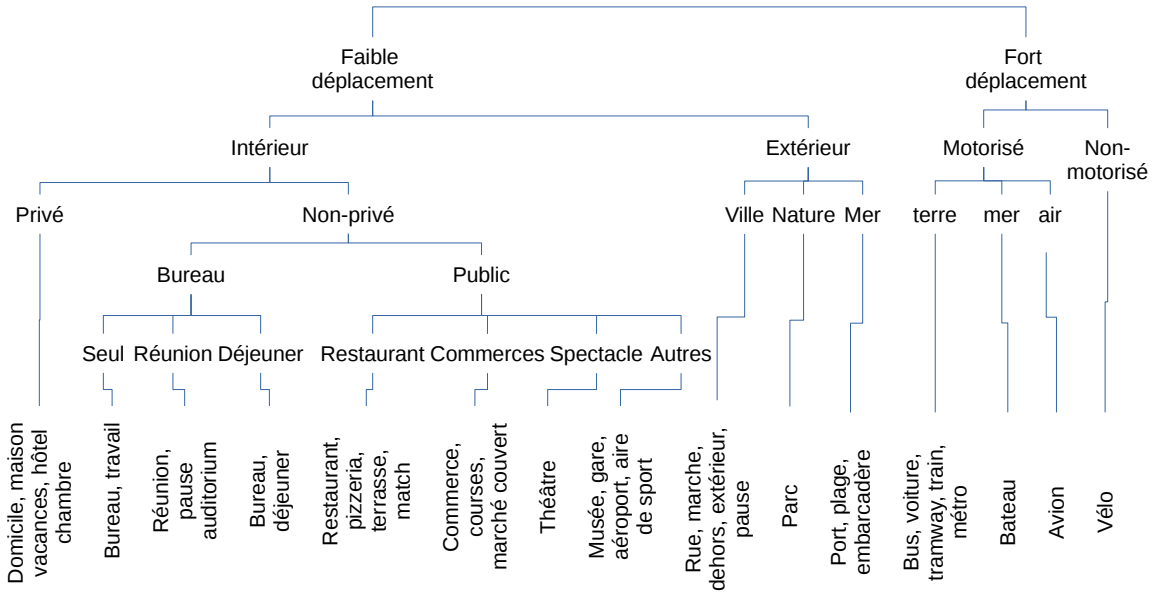


FIGURE 4.2: Proposition de regroupement des annotations collectées

4.1.2 Étude des définitions de scènes existantes

L'objectif de cette étude est de confirmer la légitimité des éléments mis en évidence précédemment ainsi que d'élargir la réflexion sur la scène.

Au théâtre, la scène est décrite comme "*chacune des subdivisions d'un acte, définie conventionnellement et correspondant généralement à l'arrivée ou au départ de personnages*" [1]. Les scènes du théâtre sont composées d'un lieu, d'une époque, de personnages

1. définition issue du site http://www.cnrtl.fr/lexicographie/scène

et d'une action dominante. Elles représentent des situations, des moments caractéristiques dans l'histoire. Dans notre collecte, les annotations obtenues représentent également des situations vécues par les volontaires, le plus souvent associées à un lieu et aussi à une action ; et qui mettent en scène au moins une personne, le détenteur du smartphone.

Au théâtre, les limites d'une scène sont bien définies et marquées par des changements dans l'un des éléments qui composent la scène. L'étude des annotations a montré la capacité des participants d'identifier ou de reconnaître une scène. Mais les délimitations de celles-ci ne sont pas précises, comme le suggèrent les séquences relevées dans la table 4.2 qui ne mentionnent pas les transitions entre scènes. Par exemple, la première séquence de la table évoque la scène de l'aéroport puis du bus. Or, le passage de l'un à l'autre requiert au minimum un déplacement, qui peut prendre plusieurs minutes. On comprend que les transitions d'une scène à l'autre ne sont pas instantanées mais progressives, contrairement à la définition issue du théâtre.

Nous avons cherché des exemples de travaux qui ont abordé le problème de reconnaissance de scènes afin d'étudier la représentation des scènes. Le défi D-CASE introduit en section 2.2.2 présente une tâche de classification de dix scènes sonores courantes : la rue bruyante, la rue calme, le parc, le marché ouvert, le bus, le métro, le restaurant, le supermarché, le bureau et la station de métro. Nous remarquons dans cette liste la similitude avec les exemples de nos annotations et le niveau de granularité choisi pour les considérer. De plus, les scènes du défi sont d'abord caractérisées par le lieu, comme le sont nos annotations. On observe également une distinction apportée à la scène de la rue, suivant l'intensité de l'ambiance sonore. Cela suggère l'imprécision potentielle de la description d'une scène par le seul lieu symbolique.

TABLE 4.3: Tables des scènes considérées par Peltonen et coll. (2002) dans leur article

Extérieur	**Véhicules**	**Domicile**
Rue	Voiture	Salon
Route	Bus	Cuisine
Nature	Train	Salle de bains
Chantier de construction	Métro	Musique
Marché	**Lieux publics**	**Réverbérant**
Parc d'attractions	Restaurant, café	Église
Bureau et lieux calmes	Pub	Gare ferroviaire
Bureau	Supermarché	Station de métro
Salle de cours ou réunion	Pause de réunion	Grande salle (type hall)
Bibliothèque	Foule en intérieur	

Un autre exemple de reconnaissance de scène sonore (*Computational Auditory Scene Recognition*, abrégé par CASR) est proposé par Peltonen et coll. (2002) dont l'article a déjà été introduit dans l'état de l'art. Les auteurs ont considéré 26 scènes différentes, rapportées dans la table 4.3 reprise de leur article. Comme pour le défi D-CASE, on retrouve des scènes similaires à celles qui sont étudiées dans nos annotations. La table présente également des groupes de scènes, dont les critères d'établissement ne sont pas mentionnés dans l'article

mais dont les noms permettent de les interpréter. Les auteurs justifient la pertinence des groupes par les scores élevés de classification. Par exemple, dans l'évaluation de la classification des vecteurs acoustiques du groupe de véhicules contre l'ensemble des autres vecteurs (il s'agit d'un problème binaire de classification), le taux de reconnaissance atteint 94,7 %. Le résultat est encourageant pour la représentation des scènes à différents niveaux de granularité.

Nous rapportons aussi les travaux de Jacquet et coll. (2004) qui se sont intéressés à la représentation de scène pour le développement d'un système qui assiste les personnes malvoyantes dans leurs déplacements. Dans l'article, les auteurs proposent une représentation par une description hiérarchique. Trois catégories d'éléments génériques sont employés pour la description : des éléments primaires (par exemple un mur ou une porte pour décrire une pièce) ; des éléments intermédiaires, composés des éléments primaires (par exemple une pièce est composée de plusieurs murs et d'une porte) ; des éléments génériques à haut niveau qui agrègent les éléments intermédiaires (par exemple, un regroupement de bureaux de travail ou un bâtiment). La notion de hiérarchie de cette représentation est pertinente pour notre problème de définition d'une scène car elle offre plusieurs niveaux de description. En outre, le modèle de représentation proposé par les auteurs repose sur la distinction de classes (les éléments génériques) et d'instances (les exemples, les éléments réels d'une scène). Ce second point est aussi intéressant pour la description de scène, car il offre la possibilité de différencier la représentation théorique de la scène des exemples qui l'illustrent.

4.1.3 Notre proposition de scènes

Dans le cadre de la thèse, et suivant les différentes observations faites précédemment, nous proposons la définition suivante pour la scène :

> **Définition 1** **Situation** : une situation de la vie quotidienne est décrite par un ensemble de conditions dont un lieu et une action ; il s'agit d'une période continue et bornée dans le temps.

> **Définition 2** **Scène :** généralisation d'un ensemble de situations qui partagent des conditions communes dont le lieu et l'action ; une scène s'exprime par la composition d'un lieu et d'une action.

Suivant la définition du CNRTL [2] , une situation est un "*ensemble de conditions physiques et morales, stables, d'une personne à un moment donné*" [3]. Les conditions communes des situations d'une même scène sont les éléments qui composent la scène. Notre définition se limite aux éléments du lieu et de l'action. Nous avons identifié le lieu car il ressort des observations des sections précédentes. Celui-ci est décrit par une étiquette symbolique. Afin de compléter les conditions de réalisation de la situation, nous ajoutons la notion d'action

2. Centre National de Ressources Textuelles et Lexicales, dont le site est accessible à l'adresse suivante : http://www.cnrtl.fr/
3. voir le site http://www.cnrtl.fr/definition/situation

que nous définissons à partir des éléments relevés dans le Compendium d'activités physiques (2011). Cet ouvrage répertorie un vaste ensemble d'actions possibles pour une personne (organisées suivant la sémantique de l'activité exercée et l'intensité de celle-ci). Bien que notre définition se limite à la composition de ces deux éléments, il n'est pas exclu qu'il puisse exister d'autres éléments extractibles des données.

La stabilité des éléments qui composent la scène est nécessaire et, lorsqu'elle n'est plus valide, un changement de situation s'opère. Un changement de situation est nécessaire pour un changement de scène. Un tel changement est temporaire et nous l'appelons *transition* car il s'agit d'une étape entre deux situations reconnues comme des scènes. Dans l'étude des annotations collectées, nous avons vu que les transitions peuvent être associées à une scène, mais il s'agit d'une interprétation, *a priori* par rapprochement des conditions dans la scène de transition à celles de la situation précédente ou de la situation à venir.

À titre d'exemple, nous avons repris les scènes de la figure 4.1 que nous avons décomposées suivant notre définition dans la table 4.4. Les lieux sont regroupés suivant trois groupes d'environnements de la figure 4.2. Les actions choisies pour la décomposition sont des groupes d'intensité d'activité physique effectuée. Le repos regroupe toutes les phases d'immobilité dans différentes postures ainsi que les activités peu intenses pratiquées dans la posture assise telles qu'un dîner ou le travail sur ordinateur. La décomposition des scènes dans les environnements et actions est basée sur une interprétation subjective. Ainsi, nous considérons qu'il est possible de pratiquer le sport dans la rue, par exemple par la course à pied.

La table illustre la complexité des scènes qui peuvent être décrites par plusieurs environnements (par exemple un restaurant peut avoir une salle principale en intérieur et une terrasse en extérieur). Elle illustre également la correspondance des scènes pour un élément : les attitudes de repos sont possibles dans toutes les scènes illustrées.

TABLE 4.4: Description des scènes considérées suivant le modèle de scènes

Scènes	**Environnement**			**Action.**		
	Intérieur	Extérieur	Transport	Repos	Marche	Sport
Restaurant, bar	✓	✓		✓	✓	
Commerce	✓			✓	✓	
Domicile	✓	✓		✓	✓	
Pause	✓			✓		
Réunion	✓			✓		
Tramway			✓	✓		
Travail	✓			✓		
Rue		✓		✓	✓	✓

4.2 Cadre expérimental

À partir des objectifs industriels, des contraintes de collecte et de notre définition de scène, nous formulons plusieurs hypothèses qui s'appliquent aux expérimentations réalisées pour répondre aux problèmes posés. Celles-ci sont décrites dans un second temps.

4.2.1 Hypothèses générales pour les expérimentations

Nous commençons par définir les scènes, lieux et actions considérés dans les expérimentations de la thèse. Par notre définition, chaque scène est décomposée en un lieu et une action. Nous énumérons les scènes considérées et les décomposons dans la table 4.5 : *domicile*, *restaurant*, *commerce*, *bureau*, *réunion*, *pause*, *rue*, *train*, *tramway*, *voiture* et *bus*. Les scènes sont choisies parmi les annotations les plus fréquentes collectées afin d'avoir le plus grand nombre d'instances de scènes. Les lieux qui composent ces scènes sont au nombre de cinq : *intérieur privé*, *intérieur public*, *lieu de travail*, *extérieur* et *transport motorisé*. Ils reprennent les premiers nœuds de l'arbre de la figure 4.2. En outre, le bureau est l'unique lieu de travail considéré et pour lequel des données sont disponibles.

TABLE 4.5: Composition en lieux et actions des scènes étudiées

Scène	**Environnement**					**Action**			
	Int. Priv.	Int. Pub.	Lieu trav.	Ext.	Transp. Mot.	Travail ordi.	Particip. Réu.	Déplcmt à pied	Repos
Domicile	✓							✓	✓
Restaurant		✓						✓	✓
Commerce		✓						✓	✓
Bureau			✓			✓			✓
Réunion			✓			✓	✓		✓
Pause			✓					✓	✓
Rue				✓				✓	✓
Train					✓				✓
Tramway					✓				✓
Voiture					✓				✓
Bus					✓				✓

Les actions considérées sont représentatives des actions principales réalisées dans les scènes précédemment citées. Les étiquettes choisies sont le résultat d'un compromis entre les annotations des scènes et les définitions du Compendium d'activités physiques (2011). On y trouve les actions suivantes : *travailler sur ordinateur*, *participer à une réunion*, *se déplacer à pied*, *être au repos*. Les deux premières actions sont spécifiques aux environnements du lieu de travail. L'action de déplacement à pied est associée aux scènes où celui-ci est fréquent comme le *domicile*, le *restaurant*, le *commerce*, la *rue* ainsi que dans la scène de *pause* où les personnes peuvent se déplacer d'un lieu à un autre. Enfin, l'action de repos englobe les attitudes et activités exercées en posture assise, debout ou allongée, autres que celles du travail sur ordinateur et de la réunion. Par exemple, le dîner au restaurant ou l'attente dans un véhicule sont considérés comme des actions de repos. En outre, les scènes du lieu de travail sont également associées à cette action lorsque la personne n'est plus impliquée dans une autre action. Par l'observation de la table, nous remarquons que notre définition de la scène est très générale. En effet, les actions de déplacement à pied et de repos sont présentes dans toutes les scènes. Cela peut rendre difficile l'identification du lieu à partir de l'identification de l'action.

Par sa nature incertaine, la transition entre deux scènes est représentée dans le corpus par une étiquette distincte "incertain". Les transitions ne chevauchent pas les scènes, aussi une transition démarre lorsque l'on n'est plus sûr d'être dans la scène. Inversement, une

transition s'arrête lorsque l'on est certain d'être dans une nouvelle scène.

Nous formulons également deux hypothèses sur le contexte du smartphone :

Hypothèse 1 Le smartphone est soit porté par l'utilisateur, soit posé à proximité immédiate de celui-ci, de sorte que les mesures effectuées puissent être assimilées à la scène de l'utilisateur.

Hypothèse 2 Le smartphone est porté par une seule personne.

La première hypothèse considère la proximité de l'appareil et de l'utilisateur pour pouvoir affirmer que les données collectées sont représentatives du contexte de l'utilisateur. Pour rappel, nous avons vu dans l'état de l'art que le smartphone dispose d'un contexte qui lui est propre, notamment marqué par sa position relativement à l'utilisateur et son orientation. Nous avons proposé une modélisation du contexte plus vaste (Blachon et coll. 2014a), qui tient compte de la présence d'interaction (décrite par une valeur binaire) et de la quantité de mouvement de l'appareil (exprimée par trois valeurs symboliques : nulle, faible à modérée, forte), en plus de la position de l'appareil (décrite par quatre valeurs : à la main, dans la poche, dans un sac ou posé sur une surface). Dans le cadre de la thèse, nous nous limitons à l'hypothèse 1 qui est plus générale.

La seconde hypothèse considère l'unicité de l'utilisateur du smartphone, qui permet de considérer que les situations mesurées par l'appareil sont vécues par une personne unique. Cette hypothèse est pertinente pour la création d'un système de reconnaissance adapté à l'utilisateur.

4.2.2 Description des expérimentations

Nous décrivons dans cette section les expérimentations effectuées pour répondre aux problèmes posés. Les deux premières expérimentations proposent deux solutions au problème de reconnaissance de scène et constituent des références de performance. Dans un second temps, nous proposons deux expérimentations plus exploratoires. D'abord, nous souhaitons compléter la réflexion sur la connaissance de la scène par l'étude non-supervisée des vecteurs, dans le but de mettre en évidence des motifs qui pourraient être identifiés. Également, nous souhaitons proposer une solution alternative au problème de reconnaissance de scène, plus flexible et plus ouverte aux améliorations que la solution par classification décrite ci-après.

Reconnaissance de scène par classification des vecteurs

Le problème de la reconnaissance de scène est abordé en considérant un système centré sur un utilisateur, suivant les scènes qu'il a vécues. Cela représente un cadre de fonctionnement réaliste pour le système final, où les modèles sont entraînés à partir des données collectées pendant l'utilisation. En outre, ce cadre permet de tirer profit du déséquilibre entre les participants dans les données collectées.

Dans ce cadre, nous proposons un système de classification des vecteurs de descripteurs. L'évaluation du système est effectuée suivant différentes combinaisons de sources de données et de descripteurs : quand l'ensemble des capteurs considérés est disponible ; lorsque seuls l'accéléromètre et le microphone fournissent des mesures ; avec les données tous les capteurs considérés, après une sélection des attributs les plus pertinents. Nous effectuons cette évaluation pour confronter et compléter les résultats de la comparaison de sources de données menée dans la section 2.3 de l'état de l'art.

L'expérimentation est menée suivant deux méthodes de validation qui représentent des fonctionnements différents du système. La première méthode est dite à validation croisée stratifiée à dix sous-ensembles (ou en anglais, *10-fold Stratified Cross-Validation*, abrégée par *FOLD-CV* dans la suite). Elle consiste en un découpage du corpus en 10 groupes contenant chacun le même nombre de vecteurs. Un groupe est arbitrairement affecté au corpus de test pour évaluer le classifieur entraîné sur l'ensemble des neuf autres. L'opération est répétée 10 fois, de sorte que chaque groupe serve une fois à l'évaluation. Les performances sont moyennées sur l'ensemble des dix répétitions. En outre, nous appliquons cette méthode en vérifiant que chaque scène est équitablement répartie dans tous les groupes. Cependant, nous ne tenons pas compte de l'instance de scène dont proviennent les vecteurs. Ainsi, deux vecteurs différents d'une même instance de scène peuvent se retrouver dans le corpus d'entraînement et dans le corpus de test. Cette configuration représente un cas de fonctionnement où le système est adapté à des données collectées dans un passé très proche.

La seconde méthode se différencie de la précédente par la répartition uniforme des vecteurs de scènes dans le corpus d'entraînement. Elle est justifiée par la volonté d'éviter les biais dus aux classes majoritaires dans le corpus d'entraînement. Contrairement à la méthode précédente, celle-ci n'est effectuée qu'une seule fois. La comparaison des mesures de classification obtenues à celles de la validation précédente permet d'évaluer l'impact du déséquilibre des classes durant l'entraînement.

Détection des transitions entre les scènes

Nous proposons une seconde approche pour la reconnaissance de scène, basée sur la détection des transitions. Elle repose sur l'hypothèse que les transitions entre scènes, définies comme des situations incertaines et associées à aucune scène connue, sont marquées par des ruptures qu'il est possible d'identifier. Par exemple, le changement de lieu peut être perçu par le changement d'ambiance sonore ou par le déplacement de la personne.

De tels changements peuvent être occasionnés lorsque la scène ne change pas, c'est pourquoi il est raisonnable de penser que le nombre de fausses détections puisse être élevé. Nous envisageons de compléter le système avec un classifieur dont le rôle est de "lisser" les prédictions de transitions afin d'éliminer les fausses prédictions. Toutefois, l'expérimentation proposée se limite à la détection de transitions et ne vise pas l'évaluation du classifieur "lisseur".

Cette approche s'oppose à la classification de vecteurs indépendants présentée précédemment car, pour la détection de ruptures, le système considère des séquences de vecteurs

consécutifs.

Approche de découverte des données

La seconde partie des expérimentations est guidée par une approche plus exploratoire. Dans un premier temps, nous proposons un travail d'interprétation des données dans le but de compléter notre proposition de modèle de scène. Pour cela, nous présentons des résultats de regroupement de données réalisé de manière non-supervisé. Les groupes obtenus sont interprétés par des hypothèses sur la composition des scènes.

Le regroupement est effectué sur les vecteurs contenant les descripteurs d'accélérations et d'ambiance sonore. Nous distinguons d'abord les deux sources de données. L'étude des groupes de vecteurs d'accélérations seuls est justifiée par l'interprétation physique des descripteurs employés (moyenne et variance d'accélération) qui peuvent être associés à des orientations du téléphone ou des quantités de mouvement. En outre, la comparaison des groupes obtenus avec les scènes permet des hypothèses complémentaires sur le sens à donner aux groupes ou sur la composition des scènes.

Les descripteurs acoustiques ne peuvent pas être évalués de manière absolue car les coefficients d'énergie des filtres sont normalisés. Cependant, il est possible de comparer les groupes entre eux ainsi qu'avec les scènes. La comparaison des groupes avec les scènes repose sur des hypothèses faites sur l'ambiance sonore des scènes. Nous étudions la vraisemblance de ces hypothèses par la mise en évidence de signatures acoustiques dans les groupes de vecteurs concernés. Ces signatures sont observées dans les histogrammes de coefficients d'énergie des groupes.

L'interprétation des groupes de vecteurs acoustiques et d'accélérations repose sur l'idée que la considération de vecteurs avec les deux sources de données mènera à un regroupement différent. Ainsi, nous comparons les groupes obtenus dans ce cas avec les groupes des deux cas précédents.

Approche par combinaison d'éléments de scènes

La dernière expérimentation présentée propose un système de reconnaissance de scène composite. Dans le système, la scène est représentée suivant la notion de composition que nous avons introduite à la section 4.1.3. Ainsi, le système cherche à reconnaître les deux éléments de lieu et d'action qui la composent pour inférer la scène la plus vraisemblable.

Ce système représente une solution à plusieurs des problèmes de la thèse. D'abord, la description de la scène reconnue par le système est plus détaillée qu'avec l'usage d'une seule étiquette, ce qui peut être un avantage pour une application industrielle, tant en terme de qualité d'information qu'en terme de gestion de capteurs. En effet, l'usage de modules intermédiaires permet d'envisager l'exploitation différenciée des sources de données. Par exemple, un module emploie l'accéléromètre et un autre le microphone. Ainsi, si les données d'une source sont manquantes, le système peut fournir une estimation de scène et une description avec les propositions des modules valides.

La réalisation d'un tel système requiert, d'une part, la création des modules intermédiaires et, d'autre part, la mise en place d'une stratégie de combinaison des estimations des modules. Pour les modules intermédiaires de reconnaissance de lieu et d'action, nous choisissons des classifieurs dont les détails d'entraînement et d'évaluation sont précisés dans la section 6.2.

La stratégie de combinaison adoptée s'appuie sur la théorie de fusion d'évidence de Dempster-Shafer (1968; 1976). Cette théorie a l'avantage de représenter la quantité d'incertitude de réalisation d'un concept, qui complète la simple probabilité de réalisation, que l'on retrouve dans l'inférence bayésienne. Dans le cadre de l'expérimentation, l'usage des classifieurs implique la représentation par des probabilités et la théorie de Dempster-Shafer peut paraître inadaptée. Cependant, cette méthode peut s'avérer pertinente dans le cadre d'un plus grand nombre de modules intermédiaires d'inférence avec, en particulier, l'intégration d'informations issues de sources événementielles sur l'usage du téléphone (par exemple l'activation de l'écran ou l'usage d'une application).

4.3 Bilan du chapitre

À partir d'une étude des annotations collectées et des définitions existantes, nous avons mis en évidence des éléments pour la caractérisation d'une scène, qui ont servi à établir notre définition de scène. Les éléments qui composent une scène sont le lieu et l'action. Cependant, l'observation des annotations a aussi montré que la description des scènes par ces deux seuls éléments peut être insuffisante. De plus, nous avons remarqué une grande variabilité dans les scènes. Cela suggère que notre définition est très générale.

Par la suite, la combinaison des objectifs industriels, des contraintes de collecte et de la définition de scène nous a amenés à formuler plusieurs considérations, en particulier sur les scènes considérées et le contexte du smartphone. Ces considérations représentent des hypothèses qui encadrent les travaux que nous présentons.

Dans les travaux que nous présentons, nous souhaitons d'abord apporter une référence de performance au problème de reconnaissance de scène, évalué suivant plusieurs méthodes et configurations de sources de données. Également, nous complétons la solution de référence avec la présentation d'un autre système, censé répondre à plusieurs des objectifs industriels visés. Enfin, nous considérons que l'imprécision dans la définition de la scène constitue un problème important, c'est pourquoi nous proposons un travail d'interprétation des données collectées afin de faire émerger des concepts de manière non-supervisée. Le chapitre suivant décrit les deux premières expérimentations tandis que le chapitre 6 présente les approches exploratoires.

CHAPITRE 5

Expérimentations de reconnaissance de scène

Le présent chapitre aborde le problème de la reconnaissance de scène par la proposition d'un système de classification de vecteurs de descripteurs annotés avec les étiquettes des scènes collectées. Un ensemble de descripteurs est choisi pour calculer des représentations des mesures issues des sources de données matérielles (les capteurs). Les descripteurs sont d'abord évalués dans le cadre d'une sélection mettant en œuvre deux méthodes basées sur des critères différents. Les classifieurs sont choisis à partir de travaux pertinents de l'état de l'art.

Les expérimentations décrites dans le chapitre visent l'évaluation de la combinaison des descripteurs et classifieurs dans différentes conditions d'utilisation réalistes. D'abord, le corpus est limité aux données d'un volontaire et permet de considérer l'étude d'un système de reconnaissance entraîné sur les données d'une seule personne. Ensuite, trois configurations de sources de données et de capteurs sont employées pour évaluer l'évolution des résultats de classification. Enfin, deux méthodes de validation sont employées pour l'évaluation. La première, dite croisée stratifiée à 10 sous-ensembles, effectue l'évaluation des classifieurs en exploitant au mieux le corpus pour l'entraînement. La seconde méthode considère un corpus d'entraînement plus petit, mais dont la distribution des scènes est la plus uniforme possible.

Le chapitre décrit aussi une autre approche, basée sur la détection de transitions. L'hypothèse sous-jacente est que les transitions sont exprimées par des changements de lieu ou d'action, qui peuvent être détectés par les ruptures dans les mesures de capteurs.

5.1 Description des ensembles de données et des classifieurs

Dans cette section, nous présentons les éléments qui ont permis les expérimentations du chapitre : le corpus, les descripteurs, les classifieurs, les outils et les mesures de performances.

5.1.1 Description du corpus

Le corpus employé dans les expérimentations de ce chapitre est constitué de 22 enregistrements, tous effectués des jours différents et par un seul volontaire. Les enregistrements sont considérés comme indépendants les uns des autres. Suivant les sources de données et

les descripteurs considérés, nous employons dans le chapitre plusieurs configurations du même corpus, que nous décrivons ci-dessous :

— la configuration *REF* fait référence à la composition de l'ensemble des descripteurs des cinq sources de données considérées ;

— la configuration *REF_SA* fait référence à une sélection des descripteurs (décrite dans la suite) provenant également des cinq sources de données ;

— la configuration *REF_AccAud* fait référence à l'ensemble des descripteurs des données de l'accéléromètre et du microphone, soit 68 descripteurs.

La composition en scènes des trois corpus est représentée dans la table 5.1. Nous reprenons le concept de situation, défini dans la section 4.1.3, pour décrire une instance de scène continue pendant laquelle le volontaire se situe dans la scène. Une situation s'arrête lorsque l'on change de scène. Ainsi, la table indique qu'il y a 8 situations de la scène du *magasin* réparties dans 5 enregistrements différents. Parmi les observations notables, la scène du *train* a une situation unique. Au total, 219 situations de scènes se répartissent dans les 22 enregistrements et correspondent à 108,8 heures de données collectées pour un volontaire unique. On remarque également la présence d'incertitudes dans 21 des 22 enregistrements. Chaque transition entre deux scènes est marquée par une incertitude, il est donc attendu d'en trouver dans tous les enregistrements. L'absence de l'un des enregistrements est expliquée par la scène unique du *bureau* dans l'un d'eux, qui n'a pas donné lieu à des changements et donc des transitions. La durée moyenne d'une incertitude est de 116 secondes (obtenue par le ratio de la durée cumulée des incertitudes sur le nombre de transitions). Enfin, les vecteurs des corpus sont normalisés en moyenne et variance.

TABLE 5.1: Détails des trois corpus employés dans les expérimentations

Scène	**Instances**	**Enreg.**	**Durée (s)**	**Durée (h)**
Train	1	1	4200	1,2
Pause	4	2	5376	1,3
Magasin	8	5	9040	2,5
Voiture	12	7	12214	3,4
Tram	26	14	14940	4,2
Bus	11	8	16367	4,5
Rue	92	21	29891	8,3
Réunion	9	7	42347	11,8
Restaurant	14	10	45874	12,7
Domicile	25	20	88547	24,6
Bureau	17	14	100204	27,8
Total	219	-	391648	108,8
Incertitude	*199*	*21*	*23235*	*6,5*

Les descripteurs sont calculés sur des fenêtres de signal glissantes longues de deux secondes et avec un recouvrement d'une seconde. Les données issues des sources de l'accéléromètre, du gyroscope, du magnétomètre et du baromètre ont donné lieu au même calcul de descripteurs. Pour chacun, la moyenne, la variance et l'énergie sont calculées à partir de

la norme du signal dans la fenêtre. Si l'on considère le vecteur $X = [x_1, \ldots, x_n]$ contenant la norme des échantillons consécutifs sélectionnés par une fenêtre, les équations ci-dessous décrivent les formules pour le calcul de ces trois descripteurs. Le reste des descripteurs provient de l'application d'une transformée de Fourier sur le signal, non-uniforme car les échantillons ne sont pas périodiques. Celle-ci a pour limite supérieure de fréquence 50 Hz. Les coefficients d'énergie dans les bandes de fréquence entre 0 et 25 Hz sont calculés et gardés comme descripteurs.

$$\begin{aligned} \mu &= \frac{1}{n}\sum_{i=1}^{n} x_i \\ \sigma &= \frac{1}{n}\sum_{i=1}^{n} (x_i - \mu)^2 \\ e &= \sum_{i=1}^{n} x_i^2 \end{aligned} \tag{5.1}$$

Le calcul des descripteurs acoustiques est représenté dans la figure 5.1. Il est effectué sur des fenêtres plus courtes de 23 millisecondes (soient 1024 échantillons de signal à une fréquence d'échantillonnage de 44,1 kHz). Une transformée de Fourier uniforme est appliquée sur la fenêtre du signal, dont on ne garde que les coefficients d'amplitude. Par la suite, un banc de 40 filtres triangulaires est appliqué aux coefficients d'amplitude. Les filtres sont régulièrement répartis sur l'échelle des Mel correspondant à l'échelle Hertz entre 0 et 22050 Hz (la fréquence de Nyquist pour le signal acoustique). La figure 5.2 illustre la répartition des filtres. La table 7.1, située en annexe de ce manuscrit, décrit les fréquences de début, centre et fin de chaque filtre. Le résultat du filtrage est un vecteur de 40 coefficients d'énergie. Enfin, les vecteurs sont moyennés sur des fenêtres de 2 secondes afin de les aligner avec les descripteurs des autres sources.

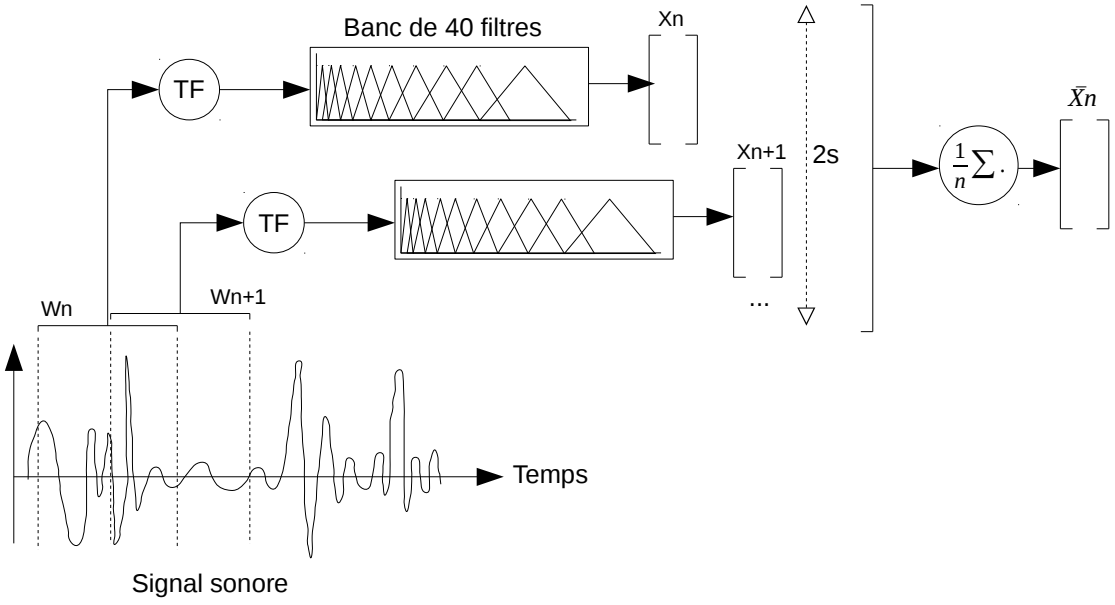


FIGURE 5.1: Schéma du calcul des descripteurs acoustiques

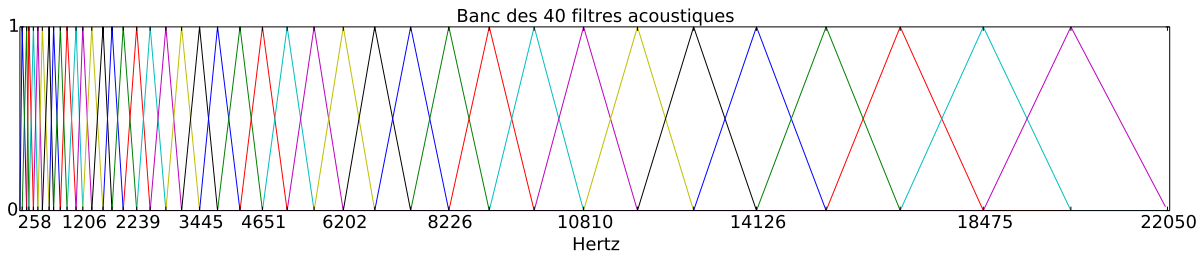


FIGURE 5.2: Répartition en fréquence Hertz des 40 filtres acoustiques

5.1.2 Description des classifieurs et mesures d'évaluation

Les classifieurs sélectionnés pour ces expérimentations sont les arbres de décision C4.5 (abrégé C4.5) et la forêt d'arbres décisionnels (abrégé RF), le classifieur à base de GMM (abrégé GMM), le réseau bayésien naïf (NB) et le réseau de neurones artificiel DNN (abrégé DNN). La théorie sous-jacente aux modèles et algorithmes, ainsi que les usages de chacun d'eux ont été introduits dans le chapitre de l'état de l'art à la section 2.4. Les arbres C4.5 et RF ont permis d'obtenir des résultats élevés de reconnaissance d'activité physiques et postures, le NB représente une référence dans de nombreux domaines, le GMM est beaucoup employé dans le domaine acoustique et le DNN a récemment donné lieu à de très bonnes performances dans la reconnaissance de locuteurs et d'images.

Nous avons utilisé l'outil Weka (Hall et coll. 2009) pour l'entraînement et l'évaluation des classifieurs C4.5, RF et NB. L'implémentation du C4.5 dans Weka est appelée J48. Nous avons généré un arbre élagué établi par la méthode de réduction d'erreur par corpus de validation. Ainsi, un sous-ensemble du corpus est dédié à la validation de l'arbre pendant l'entraînement. Une autre mesure prise pour limiter le sur-apprentissage de l'arbre a consisté à limiter la taille minimale des feuilles à 100 vecteurs. L'entraînement du Random Forest a été effectué en calculant 100 arbres de décision. L'entraînement du NB a utilisé des distributions normales pour l'estimation des probabilités des attributs numériques.

Le classifieur à GMM a été entraîné et évalué *via* la bibliothèque de code *alize* (Larcher et coll. 2013). Un modèle du monde est d'abord créé à partir de l'ensemble des vecteurs du corpus d'entraînement ainsi que des vecteurs de transitions (portant l'étiquette "incertain"). Le modèle du monde GMM est composé de 512 distributions gaussiennes initialisées par un processus de 15 itérations. Par la suite, les modèles de chaque scène sont adaptés du modèle du monde par la méthode du "maximum a posteriori". Pour chaque vecteur de test, l'évaluation consiste à calculer la vraisemblance que le modèle ait généré le vecteur (ou la probabilité *a posteriori* du modèle connaissant le vecteur). La vraisemblance retournée correspond à un score. Finalement, deux normalisations successives sont appliquées pour comparer les scores des différents modèles pour un vecteur de test. La première consiste à évaluer l'échantillon par le modèle du monde et à retirer le "score" du modèle du monde (la même probabilité *a posteriori* du modèle que pour le modèle de scène) au score du modèle de scène. La seconde est une normalisation en moyenne et variance sur les scores d'un modèle afin de centrer et réduire la distribution des scores d'un modèle pour ensuite permettre

la comparaison des scores entre modèles.

L'utilisation du DNN repose sur la bibliothèque de code *PDNN* (Miao 2014) pour le langage Python. Le DNN utilisé est un réseau de neurones à structure orientée (en anglais, *feed-forward neural network*) composé d'une seule couche cachée dont le nombre de nœuds est égale à la moitié du nombre d'attributs en entrée. La fonction sigmoïde est employée pour activer les neurones. L'évolution du taux d'apprentissage est progressive pendant l'entraînement. Pendant les 50 premières itérations, il est fixé à une valeur constante de 0,08. Par la suite, la différence d'erreur sur le corpus de validation entre deux itérations successives est évaluée et si elle est inférieure à 0,05, le taux d'apprentissage est divisé par 2. Finalement, si cette même différence se maintient en-dessous de 0,05 alors l'entraînement s'arrête.

Enfin, nous définissons les mesures employées au cours de ce chapitre pour exprimer les résultats.

Prédictivité positive La *prédictivité positive* représente la capacité du système à correctement identifier les échantillons d'une classe. Elle s'exprime par la formule suivante :

$$\text{prédictivité} = \frac{TP}{TP+FP} \tag{5.2}$$

Dans la suite du chapitre, nous employons le terme "précision" pour évoquer la prédictivité positive.

Rappel Le *rappel* (appelé également la sensibilité) mesure la capacité du système à retrouver tous les échantillons d'une classe et s'exprime par la formule suivante :

$$\text{rappel} = \frac{TP}{TP+FN} \tag{5.3}$$

Ces définitions peuvent être généralisées lorsque le nombre de classes est supérieur à 2. Pour la précision d'une classe C_i, le dénominateur exprime l'ensemble des prédictions du classifieur associées à cette classe. Pour le rappel, le dénominateur exprime l'ensemble des échantillons qui vérifie la classe C_i.

F-Mesure La *F-Mesure* exprime une combinaison du rappel et de la prédictivité positive en un seul score. Sa formule est rappelée ci-dessous.

$$F = 2.\frac{\text{précision.rappel}}{\text{précision} + \text{rappel}} \tag{5.4}$$

Précision et rappel moyens Les précision et rappel moyens sont calculés en moyennant les précision et rappel des classes individuelles par le nombre de classes.

Taux de classification Ce taux représente le nombre d'échantillons correctement reconnus sur le nombre total d'échantillons évalués.

5.2 Sélection d'attributs

Les descripteurs introduits dans la section 5.1.1 sont issus de l'étude de travaux du chapitre 2, en particulier de ceux mettant en œuvre l'accéléromètre, le gyroscope et le microphone. Les sources du magnétomètre et du baromètre ne sont pas mentionnées dans ces travaux, aussi nous avons repris les mêmes descripteurs que pour l'accéléromètre et le gyroscope.

Dans la section 2.4.1, nous avons écrit que la sélection d'attributs vise à réduire l'impact des attributs non-pertinents en mettant en avant ceux qui le sont. Dans notre cas, la sélection vise aussi à évaluer la pertinence des sources de données. En particulier, cela représente un moyen d'évaluer la pertinence de l'accéléromètre et du microphone, apparus comme deux sources incontournables pour les traitements sur le smartphone après l'étude de l'état de l'art.

5.2.1 La sélection d'attributs par ratio de gain d'information

La sélection par ratio de gain d'information représente une mesure de sélection objective de descripteurs où l'information d'une scène est mesurée par la différence d'entropie de la scène *a priori* et lorsque l'on connaît une autre variable (un descripteur) [1]. Les descripteurs sont évalués individuellement et peuvent être classés suivant le gain d'information. Cette mesure est pertinente pour notre problème où le concept de scène n'est pas bien défini.

Nous avons utilisé l'implémentation du logiciel Weka (Hall et coll. 2009) pour la réalisation. Les valeurs de ratio de gain d'information obtenues varient entre 0,171 pour le descripteur *moyenne pression* et 0,004 pour le descripteur *Bande 17 Hz pression*. Nous présentons dans un premier temps la répartition des rangs des descripteurs suivant les sources de données dans la figure 5.3. L'axe des ordonnées indique le rang de l'attribut : le rang 1 étant le plus élevé et donc celui de l'attribut le plus informatif. Pour chaque source, une "boîte" représente l'étendue entre les premier et troisième quartile et un segment horizontal dans la boîte représente le rang médian. En dehors de la boîte, les attributs sont représentés par des points.

La première observation à faire sur la figure porte sur les distributions des descripteurs de pression et de champ magnétique, toutes les deux compactes et concentrés dans les rangs les plus bas. L'exception est le descripteur de la moyenne de la pression, qui se situe au premier rang et représente donc le descripteur le plus informatif.

Le gyroscope et l'audio sont les deux sources les plus informatives suivant la figure. Les rangs médians des deux distributions sont en-dessous de 50 et très proches l'un de l'autre. La répartition des descripteurs du gyroscope est plus compacte que celle des descripteurs acoustiques qui s'étend des premiers rangs au-delà du centième. Les descripteurs de l'accéléromètre occupent une position intermédiaire, centrés autour d'un rang médian proche de 70 et absents des 20 premiers rangs (le premier descripteur d'accélération est la bande

1. Nous renvoyons le lecteur à la section 2.4.1.2 pour la description de la mesure

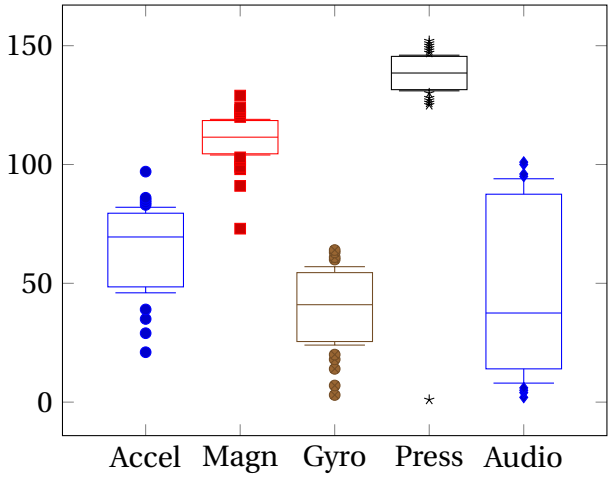


FIGURE 5.3: Distribution des rangs des attributs suivant les sources de données suite à la sélection par ratio de gain d'information

d'énergie à 3 Hz, classé 21$^{\text{ième}}$).

L'étude plus précise du classement des descripteurs (voir la table 7.2 en annexe) indique que les coefficients acoustiques de la première moitié de bande (indices 0 à 20) figurent parmi les 40 premiers rangs. Les coefficients d'indices supérieurs à 30 sont classés autour du centième rang. La variance de mesure du gyroscope est le coefficient le plus élevé de cette source (troisième rang). Viennent ensuite les énergies dans les bandes de fréquence de 1 à 10 Hz (rangs 18 à 31). Les descripteurs les mieux classés pour l'accéléromètre sont également les bandes d'énergie entre 1 et 10 Hz, ainsi que la variance d'accélération (29$^{\text{ième}}$) et l'énergie (35$^{\text{ième}}$). Le premier descripteur du champ magnétique est la moyenne du champ, classée 73$^{\text{ième}}$.

Les résultats de la sélection par corrélation

L'évaluation individuelle précédente ne tient pas compte de la combinaison des descripteurs qui peut parfois ajouter de la redondance ou dégrader la performance. C'est pourquoi nous employons une méthode de sélection par ensemble d'attributs. Le critère d'évaluation, appelé "mérite" (défini dans la section 2.4.1.2) calcule le rapport de la pertinence du sous-ensemble de descripteurs (par la corrélation aux classes) divisé par la redondance intrinsèque (exprimée par l'inter-corrélation entre les descripteurs). Ainsi, le critère pénalise les sous-ensembles peu pertinents ou très redondants. Le critère est combiné avec la méthode de génération heuristique de sous-ensembles appelée *Sequential forward selection* qui démarre avec un ensemble vide et ajoute progressivement le descripteur qui augmente le plus le mérite du sous-ensemble. L'ensemble de la méthode ne permet d'atteindre qu'un sous-ensemble optimal mais procède rapidement (voir la section 2.4.1.1).

La figure 5.4 illustre l'évolution du mérite à mesure que les descripteurs sont ajoutés. On observe une augmentation rapide du mérite, puis un plateau, qui indique la valeur maximale du mérite et le sous-ensembles optimal. La table 5.2 détaille le classement des descripteurs jusqu'au plateau.

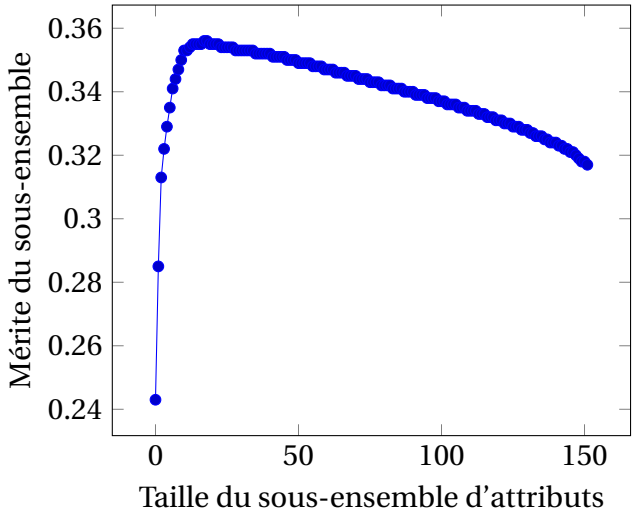


FIGURE 5.4: Distribution des scores des sous-ensembles d'attributs sélectionnés suivant l'algorithme CFS et une méthode d'évaluation heuristique progressive

TABLE 5.2: Classement des attributs dans le sous-ensemble qui donne le meilleur score suivant la méthode de corrélation

Rang	Score	Attribut	Rang	Score	Attribut	Rang	Score	Attribut
1	0.243	Moyenne pression	8	0.344	Variance accél.	14	0.355	Bande 23 Hz accél.
2	0.285	Variance gyroscope	9	0.347	Variance magné.	-	-	Bande 1 Hz gyro.
3	0.313	Bande audio #1	10	0.35	Bande audio #0	-	-	Bande audio #9
4	0.322	Bande 3 Hz accél.	11	0.353	Moyenne accél.	-	-	Bande 2 Hz magn.
5	0.329	Moyenne magné.	-	0.353	Bande 3 Hz gyro.	18	0.356	Bande 4 Hz accél.
6	0.335	Bande audio #6	13	0.354	Bande audio #2	-	-	Bande 24 Hz magné.
7	0.341	Bande 2 Hz gyro.						

L'attribut de moyenne de pression est le premier descripteur à être ajouté, c'est donc celui qui est le plus corrélé aux différentes scènes. La variance du gyroscope est le second descripteur. Cinq descripteurs acoustiques sont présents dans la table, tous d'indices inférieurs à 10. Quatre descripteurs du gyroscope et cinq descripteurs de l'accéléromètre sont également présents, représentant la moyenne, variance et les coefficients d'énergie des très basses bandes de fréquence (1, 2 et 3 Hz), ainsi que l'énergie de la bande à 23 Hz pour l'accéléromètre. Les descripteurs magnétiques sont au nombre de 4, avec la moyenne, la variance et deux coefficients d'énergie de bande.

Bilan de la sélection d'attributs

Les résultats des deux méthodes présentent certains points communs. D'abord, la moyenne de la pression atmosphérique est un descripteur important. Notre hypothèse est que les scènes en environnement intérieur sont caractérisées par un lieu en étage, ce qui correspond à une légère augmentation d'altitude donc à une légère diminution de la pression atmosphérique, que les baromètres des smartphones peuvent détecter. La variation de la pression atmosphérique dépend également des conditions météorologiques. Mais dans le cas de la routine d'un utilisateur et d'une période courte dans l'année, il est possible d'imaginer des valeurs assez stables, dont les changements correspondent effectivement à des

transitions d'étages ou de lieux.

Exception faite de la moyenne de la pression, les sources du magnétomètre et du baromètre se montrent d'un usage limité, à part pour des descripteurs statistiques communs comme la variance et la moyenne. En comparaison, les trois autres sources sont plus pertinentes. Les descripteurs acoustiques les plus pertinents sont concentrés dans les filtres d'indices bas, vraisemblablement entre 0 et 10, ce qui correspond à une bande de fréquence comprise entre 0 et 1077 Hz. Les descripteurs d'accéléromètre et du gyroscope les plus pertinents sont la moyenne, la variance et des coefficients d'énergie dans les bandes de très basse fréquence (moins de 5 Hz).

Les descripteurs retenus pour la configuration *REF_SA* sont les suivants :

— accéléromètre et gyroscope : moyenne, variance, énergie et coefficients d'énergie de la transformée de Fourier de 0 à 5 Hz ;

— baromètre : moyenne ;

— magnétomètre : moyenne et variance ;

— microphone : les 10 premiers coefficients d'énergie (représentant l'énergie dans la bande de fréquences équivalente de $[0; 1077 Hz]$

5.3 Résultats expérimentaux en validation croisée

Dans la section, nous présentons les résultats de l'expérimentation de classification de vecteurs de scène. Celle-ci est évaluée suivant une méthode de validation croisée stratifiée à 10 sous-ensembles et dans les trois configurations de capteurs introduites à la section 5.1.1 (*REF*, *REF_SA* et *REF_AccAud*).

5.3.1 Corpus *REF*

La table 5.3 résume les performances moyennes de reconnaissance des différents classifieurs sur les dix sous-ensembles du corpus de la validation croisée. Les classifieurs les plus performants sont l'arbre de décision C4.5 et la forêt d'arbres décisionnels (RF), avec des taux de classification très proches, autour de 90 % (pour rappel, le taux de classification défini dans la section 5.1.2 indique le pourcentage de vecteurs correctement identifiés). Viennent ensuite le DNN et le GMM avec des taux respectivement de 68,9 % et 40,8 %. Enfin, le réseau bayésien naïf NB affiche 10,5 %, à peine plus que la probabilité de tirer aléatoirement la bonne classe parmi les 11 présentes dans le corpus *REF*.

TABLE 5.3: Taux de classification moyen et écart-type calculés sur les 10 sous-ensembles de la validation croisée, pour la configuration de corpus *REF*

	C4.5	**RF**	**NB**	**GMM**	**DNN**	*Hasard*
Taux de classification	89,7 ± 0,2	90,3 ± 0,2	10,5 ± 0,5	40,8 ± 1,8	68,9 ± 0,6	*9,1*

Nous formulons plusieurs hypothèses pour expliquer ces tendances de performance. D'abord, il faut remarquer que le découpage du corpus en dix sous-ensembles ne distingue

ni les enregistrements ni les situations d'origine. Ainsi, différents vecteurs d'un même enregistrement et d'une même situation peuvent être placés dans les corpus d'entraînement et de test. Nous pensons que cela a un effet sur les capacités de distinction des classifieurs à arbres C4.5 et RF, dont l'algorithme d'entraînement vise à diviser le corpus en sous-ensembles homogènes et identifiables par une succession de tests. En effet, si des vecteurs d'une même situation se retrouvent dans les deux corpus d'entraînement et d'évaluation, certains des tests appris pendant l'entraînement peuvent correspondre à des vecteurs du corpus d'évaluation.

Le fonctionnement du NB repose sur l'hypothèse de la dépendance conditionnelle directe de la classe aux variables d'entrée, c'est-à-dire les descripteurs des vecteurs. Nous pensons que cette hypothèse est incorrecte et notre définition de scène de la section 4.1.3 suggère une composition avec des éléments intermédiaires. Les faibles scores du NB, proches des valeurs obtenues par tirage aléatoire, confortent l'idée d'absence de dépendance conditionnelle directe.

Les deux classifieurs GMM et DNN sont réputés nécessiter un grand nombre de données pour l'entraînement. Nous pensons que la composition du corpus est insuffisante pour ces modèles. En outre, ces modèles ont des paramètres qu'il est nécessaire d'adapter pour améliorer la reconnaissance. Malgré cela, le DNN présente des valeurs de rappel proches de 80 % pour certaines classes comme le *bus*, la *voiture*, le *bureau*, le *train* ou le *tramway* (voir table 5.5). Ainsi, il peut y avoir une autre explication aux performances médiocres de ces deux classifieurs. Nous étudions les matrices de confusion de ces deux classifieurs (tables 5.4 et 5.5) pour tenter de compléter l'explication.

Les matrices présentées sont le résultat de la somme des matrices de confusion obtenues sur les dix sous-ensembles de test de la validation croisée stratifiée, normalisées en rappel (soit par le nombre de vecteurs à retrouver dans chaque scène). Dans la table 5.5 de la matrice du DNN, les rappels des scènes du *domicile*, de la *réunion*, de la *pause*, du *restaurant* et du *magasin* sont inférieurs à 70 % et descendent parfois très bas (seulement 5,3 % pour la pause). De plus, ces scènes présentent des confusions entre elles. Concernant le GMM, les scores de rappel des scènes sont globalement plus bas (voir la table 5.4). Aux confusions observées pour le DNN, s'ajoutent celles avec le *bus*, la *voiture*, le *bureau* et la *rue*.

TABLE 5.4: Matrice de confusion du GMM

a	b	c	d	e	f	g	h	i	j	k	← Reconnu
33.2	7.2	2.8	**22.7**	1.5	3.0	4.7	**8.9**	3.7	6.7	5.6	a = Bus
4.7	**57.6**	0.7	**12.3**	3.6	1.0	0.7	3.0	0.8	**13.0**	2.5	b = Voiture
0.6	0.8	**24.0**	**27.2**	7.6	**26.5**	7.0	2.7	2.8	0.5	0.3	c = Domicile
0.6	0.6	**16.3**	**46.4**	8.2	**19.7**	2.3	2.8	2.3	0.3	0.4	d = Réunion
1.4	1.1	8.7	2.7	**40.2**	**34.3**	2.2	4.7	3.5	0.2	0.9	e = Pause
0.3	0.6	**14.2**	**12.9**	5.6	**56.3**	1.2	1.0	1.2	6.6	0.1	f = Bureau
1.2	1.6	**13.0**	5.3	**11.4**	**28.7**	**19.8**	**9.9**	7.0	1.6	0.6	g = Restaurant
2.6	1.6	**9.8**	7.3	2.3	8.6	**13.5**	**40.3**	6.1	5.7	2.2	h = Magasin
3.1	1.6	**7.8**	**8.2**	2.0	**8.3**	**8.2**	**8.0**	**49.0**	1.8	1.9	i = Rue
0.8	4.5	0.7	0.0	0.1	**11.0**	0.0	1.8	0.4	**75.2**	5.6	j = Train
2.6	1.1	5.3	**8.5**	1.2	6.8	4.3	5.3	2.1	**10.3**	**52.7**	k = Tramway

TABLE 5.5: Matrice de confusion du DNN

a	b	c	d	e	f	g	h	i	j	k	← Reconnu
76.8	2.7	2.2	1.7	0.0	5.8	1.4	2.8	2.1	0.0	4.5	a = Bus
5.2	**78.5**	2.7	1.9	0.0	3.4	3.8	0.4	0.3	2.4	1.3	b = Voiture
0.5	0.2	**67.5**	3.3	0.0	**20.4**	6.5	0.3	1.2	0.0	0.2	c = Domicile
0.4	0.2	**26.6**	**43.8**	0.0	**21.0**	5.1	0.3	2.1	0.0	0.6	d = Réunion
0.6	0.1	**15.4**	7.3	**5.3**	**54.1**	**9.0**	0.9	5.9	0.1	1.1	e = Pause
0.1	0.1	**8.7**	4.0	0.0	**84.0**	2.2	0.1	0.7	0.0	0.2	f = Bureau
0.3	0.2	**17.3**	3.6	0.1	**11.8**	**63.5**	0.4	2.0	0.1	0.7	g = Restaurant
3.9	0.2	**18.4**	6.1	0.0	7.5	8.0	**43.6**	**9.1**	0.0	3.2	h = Magasin
1.8	0.5	**8.4**	2.3	0.0	5.1	**7.0**	2.4	**70.7**	0.1	1.8	i = Rue
0.1	5.2	**7.8**	0.2	0.0	4.5	0.9	0.1	0.2	**79.0**	2.0	j = Train
5.3	0.4	3.8	1.8	0.0	4.4	4.0	1.2	2.7	0.4	**76.0**	k = Tramway

Contrairement aux classifieurs C4.5 et RF, l'entraînement des GMM et DNN vise à spécifier les valeurs des paramètres d'un modèle de représentation des scènes. La confusion des matrices laisse imaginer que certaines scènes ont des représentations très proches suivant les descripteurs employés. Intuitivement, on imagine la variabilité qu'il peut y avoir dans les situations d'une même scène. Par exemple, la perception d'un dîner au restaurant dépend fortement du lieu et du cadre (ambiance lumineuse et sonore par exemple). Pourtant, deux situations de dîner portent la même étiquette dans le corpus. Également, puisque la scène s'exprime, entre autres, par les actions effectuées par le porteur du smartphone, une conversation au *domicile* et une *réunion* au bureau ont probablement des points communs dans leurs représentations, mais portent deux étiquettes différentes dans le corpus. Ainsi, nous suggérons que la complexité des scènes à représenter ainsi que la quantité de données limitée explique probablement les performances limitées des classifieurs GMM et DNN.

5.3.2 Comparaison des performances suivant les capteurs employés

La table 5.6 résume les performances moyennes sur les dix sous-ensembles de test de la validation et suivant les trois configurations de capteurs étudiées (les descripteurs sélectionnés pour la configuration *REF_SA* sont issus des conclusions de la section 5.2). Les tendances de rang et d'écart de performances suivant les classifieurs sont identiques dans les trois configurations. Par ailleurs, on note une baisse généralisée des performances après la sélection d'attributs, sauf pour le RF (et le NB qui est déjà très bas). Cette observation peut être expliquée par la possible perte d'information due au retrait de certains descripteurs. La stabilité de résultat de la forêt d'arbres décisionnels (RF) peut provenir de l'entraînement particulier qui repose sur plusieurs sous-ensembles de descripteurs, ce qui peut réduire l'impact de l'absence de certains d'entre eux. La configuration *REF_AccAud* composée des seuls descripteurs d'accélération et acoustiques montre des performances encore diminuées. Toutefois, les scores du C4.5 et du RF restent très honorables.

TABLE 5.6: Taux de classification moyen et écart-type sur dans les trois configurations de capteurs, en validation croisée à dix sous-ensembles

	C4.5	**RF**	**NB**	**GMM**	**DNN**	*Hasard*
Configuration *REF_SA*	83,1 ± 0,1	91,0 ± 0,1	12,5 ± 0,1	31,4 ± 2,2	61,9 ± 0,1	
Configuration *REF_AccAud*	71,6 ± 0,1	77,2 ± 0,1	12,1 ± 0,1	37,0 ± 0,4	53,9 ± 0,3	*9,1*
Configuration REF	*89,7 ± 0,2*	*90,3 ± 0,2*	*10,5 ± 0,5*	*40,8 ± 1,8*	*68,9 ± 0,6*	

5.4 Résultats sur corpus d'entraînement uniforme

Nous présentons dans la table 5.7 la répartition des scènes dans les corpus de test et d'entraînement (abrégé *Train* dans la table). Les valeurs indiquent le nombre de vecteurs pour chaque scène dans les deux corpus, qui correspondent également à la durée cumulée en secondes. La répartition est effectuée afin d'obtenir un corpus d'entraînement uniforme pour toutes les scènes. Le nombre de vecteurs à y intégrer est contraint par la disproportion des vecteurs suivant les scènes, la volonté d'avoir un corpus d'entraînement le plus uniforme possible et un corpus de test avec un seuil minimal d'échantillons par scène. Ainsi, le nombre de vecteurs par scène à intégrer au corpus d'entraînement est fixé suivant les scènes les moins représentées, le *train* et la *pause*. Le nombre de 5000 est adopté pour toutes les scènes. La scène du *train* est très faible, aussi nous avons décidé d'adopter une répartition de 80 % des vecteurs pour le corpus d'entraînement et de 20 % pour le test, qui correspondent respectivement aux valeurs de 3360 et 840.

TABLE 5.7: Répartition des échantillons pour un corpus d'entraînement équilibré suivant les classes

	Affectation	
Scène	**Train (sec.)**	**Test (sec.)**
Train	3360	840
Pause	5000	376
Magasin	5000	4040
Voiture	5000	7214
Tram	5000	9940
Bus	5000	11367
Rue	5000	24891
Réunion	5000	37347
Restaurant	5000	40874
Domicile	5000	83547
Bureau	5000	95204

5.4.1 Classification sur le corpus *REF*

Nous présentons dans la table 5.8 les performances des cinq classifieurs entraînés avec le corpus d'entraînement uniforme, suivant la configuration de capteurs *REF*. En comparaison avec les valeurs de la table 5.3 pour la classification en validation croisée stratifiée à 10 sous-ensembles, les performances ont diminué. Le C4.5, le RF et le GMM perdent entre 4 et 10

points. La baisse la plus spectaculaire est pour le DNN qui passe de 68,9 % de reconnaissance à seulement 33,7 %.

TABLE 5.8: Mesures de performance pour le corpus *REF*

	C4.5	**RF**	**NB**	**GMM**	**DNN**
F-mesure moy.	68,4	**71,4**	13,1	34,2	36,7
Rappel moy.	81,6	**87,3**	20,0	45,9	55,0
Précision moy.	64,8	**67,2**	28,5	38,7	38,0
Taux de classification	79,4	**83,0**	10,1	36,7	33,7

La tendance générale de baisse des performances est vraisemblablement due à la forte réduction du corpus d'entraînement passant de 352000 vecteurs dans la validation croisée à 10 sous-ensembles à seulement 53000 pour l'évaluation courante, soit une division de la taille par 6. Cette forte réduction pourrait expliquer la baisse de performance importante relevée pour le DNN. De plus, les vecteurs à intégrer dans le corpus d'entraînement sont choisis aléatoirement sans contrainte sur les situations ou enregistrements dont ils proviennent. Ainsi, il est raisonnable de penser que le corpus d'entraînement contient des exemples d'une grande partie des 219 situations du total. Dans ce contexte de corpus limité en taille et représentatif d'un nombre d'exemples important, nous suggérons que le processus d'entraînement des arbres est plus adapté que celui de la génération de modèles. Il est évident que l'estimation de distributions de probabilités nécessite un nombre important d'éléments. Inversement, un nombre limité d'instances peut conduire à l'élaboration de critères pour une division efficace.

Comme pour l'expérimentation de la section précédente, nous nous intéressons aux confusions de différents classifieurs. Nous nous concentrons sur le RF qui présente le meilleur taux de reconnaissance et sur le DNN qui a subi la plus forte baisse. La table 5.9 présente la matrice de confusion du RF. Les valeurs sont rapportées sur le nombre de vecteurs de la scène à reconnaître dans le corpus de test. Par exemple, dans la première ligne de la matrice, 92,4 % des 11367 vecteurs de la scène du bus sont reconnus comme cette même scène. La plupart des valeurs de rappel approchent ou dépassent les 90 %. Les valeurs les plus faibles sont autour de 80 % et sont associées aux scènes *domicile*, *réunion*, *bureau* et *rue*. Les trois premières scènes présentent des confusions entre elles et avec le *restaurant*. La scène de *rue* est également confondue avec le *restaurant* ainsi qu'avec le *magasin*. De manière générale, on retrouve les confusions observées dans l'étude des matrices de confusion des tables 5.4 et 5.5.

Nous décrivons également la figure 5.5 qui présente les matrices de confusion du RF et du DNN sous forme graphique. Les valeurs de la matrice sont converties en niveaux de gris : plus la valeur est élevée, plus la case est sombre. Les valeurs extrêmes de 0 et 100 correspondent respectivement à des cases blanche et noire. La matrice de la table 5.9 est représentée par la matrice de gauche de la figure 5.5. On retrouve une diagonale très sombre et très marquée. En comparaison, la matrice du DNN est affichée à sa droite. Les cases de la diagonale sont globalement moins sombres et presque blanches pour certaines d'entre elles.

TABLE 5.9: Matrice de confusion des classes pour le classifieur RF exprimées en scores de rappel

a	b	c	d	e	f	g	h	i	j	k	← Reconnu
92.4	0.0	0.0	1.9	0.0	0.6	0.1	0.1	0.0	4.4	0.5	a = Bus
4.6	**91.9**	0.7	0.5	0.0	0.0	0.0	0.0	1.4	0.3	0.5	b = Voiture
0.2	2.5	**94.6**	1.0	1.1	0.0	0.1	0.1	0.1	0.0	0.2	c = Train
3.2	1.2	0.6	**88.8**	0.2	0.3	0.1	0.1	2.4	2.0	1.1	d = Tram
0.4	1.0	0.3	0.3	**83.7**	1.3	2.5	0.9	5.4	2.4	1.8	e = Domicile
0.9	0.4	0.1	0.9	3.5	**78.2**	1.2	1.3	7.5	3.2	2.8	f = Réunion
0.8	0.3	0.5	0.5	0.3	0.5	**89.9**	0.8	3.2	1.1	2.1	g = Pause
0.3	0.3	0.4	0.2	7.4	5.0	2.5	**80.1**	2.6	0.3	0.8	h = Bureau
0.8	1.0	0.1	1.0	1.6	1.8	2.1	0.2	**87.5**	1.3	2.5	i = Restaurant
1.7	0.1	0.0	0.7	0.2	0.2	0.0	0.0	3.6	**92.4**	1.1	j = Magasin
1.9	1.2	0.1	2.3	0.8	0.3	0.7	0.0	4.8	7.6	**80.3**	k = Rue

Plus précisément, les cases de la diagonale d'indices e et g (le *domicile* et la *pause*) sont très claires. Pour la ligne e, les cases grisées indiquent les fortes confusions avec les colonnes h et i (*bureau* et *restaurant*). Plus généralement, on remarque des cases légèrement grisées entre les colonnes f et j et les lignes e à j, qui correspondent aux scènes en environnement en intérieur. La scène de la *rue* présente des confusions avec le *magasin* (case de coordonnées (j,k)). Enfin, les scène des lignes h à k présentent des confusions avec le tramway en colonne d.

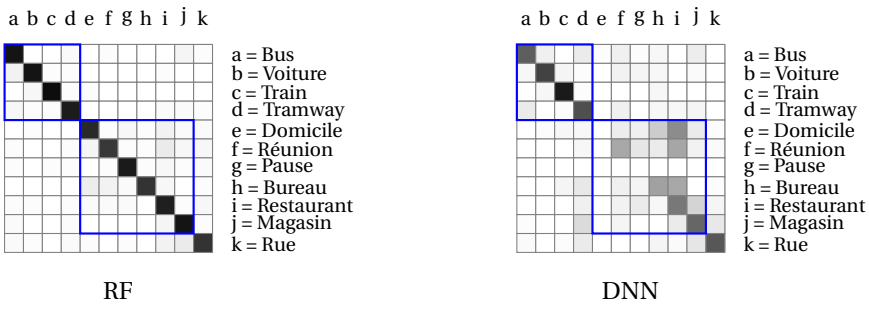


FIGURE 5.5: Matrices de confusion de reconnaissance de scènes des classifieurs RF et DNN

Dans les matrices de la figure 5.5 nous avons regroupé les scènes par macro-environnement : transports, intérieur et extérieur. Un carré de côté a à d identifie les scènes des transports ; un autre de côté e à j les scènes en intérieur ; enfin la rue est l'unique représentant en extérieur, de coordonnées (k,k). Cette représentation met en évidence que les scènes en intérieur sont essentiellement confondues entre elles.

Pour aller plus loin, nous avons recalculé les matrices de confusion suivant les macro-environnements. La table 5.6 présente les deux matrices du RF et du DNN. Les scores de rappel pour chaque environnement sont très élevés, à l'exception de l'environnement extérieur pour le DNN. Par ailleurs, la confusion entre *transports* et environnement *extérieur* semble la moins marquée. À l'inverse les environnements en *intérieur* sont plus souvent confondus avec les deux autres types de macro-environnement.

Taille		RF a	RF b	RF c	RF ← Reconnu	DNN a	DNN b	DNN c	DNN ← Reconnu
Transports	29361	**95,0**	4,3	0,7	a = Transports	**81,1**	18,3	0,6	a = Transports
Intérieur	284036	1,9	**96,5**	1,6	b = Intérieur	10,7	**87,1**	2,2	b = Intérieur
Extérieur	24891	5,5	14,2	**80,3**	c = Extérieur	13,4	19,9	**66,8**	c = Extérieur
			90,6		Moyenne		78,3		Moyenne

FIGURE 5.6: Matrices de confusion de la reconnaissance des scènes recalculée pour les groupes de macro-environnements

5.4.2 Classification sur le corpus *REF_SA*

La table 5.10 présente les performances globales des classifieurs pour l'expérimentation sur le corpus équilibré après sélection d'attributs. Pour rappel, les attributs sélectionnés sont :

— pour l'accéléromètre et le gyroscope : moyenne, variance et énergie de la norme ; énergie dans les bandes de 1 à 5 Hz ;

— pour le baromètre : moyenne ;

— pour le magnétomètre : moyenne et variance ;

— pour l'audio : les coefficients d'énergie des 10 premiers bancs de filtre sur une échelle Mel, dont l'équivalence en Hertz s'étend entre 0 et 1077 Hz (voir la table 7.1, décrite en annexe).

En comparaison avec la table 5.8, les résultats se sont globalement améliorés. Le C4.5 perd 3 points environ et le NB stagne. Le GMM et le DNN gagnent respectivement 12 et 10 points.

TABLE 5.10: Mesures de performance pour la reconnaissance de scènes après sélection d'attributs

	C4.5	**RF**	**NB**	**GMM**	**DNN**
F-mesure moy.	64,0	**82,3**	18,5	44,3	45,1
Rappel moy.	81,7	**93,7**	24,2	58,2	61,5
Précision moy.	59,4	**78,7**	34,0	43,8	45,1
Taux de classification	76,6	**92,0**	10,9	48,8	43,2

Comme précédemment, nous nous intéressons aux confusions des classifieurs. Le rappel moyen du RF est de 93,7 %, aussi nous ne présentons pas en détail sa matrice de confusion. Cependant, nous décrivons dans la figure 5.7 l'évolution des confusions du RF et du DNN.

Les matrices de la figure correspondent à la différence de la matrice de confusion pour le corpus *REF_SA* relativement à la matrice de confusion pour le corpus *REF*. Une différence nulle correspond à une case grise équilibrée (autant de noir que de blanc). Une différence positive (valeur d'une case plus élevée dans la matrice de configuration *REF_SA* que dans la matrice de configuration *REF*) est une case assombrie (la quantité de noir est augmentée et celle de blanc diminuée). Les cases assombries sur la diagonale indiquent une meilleure reconnaissance (c'est le cas du RF). Inversement, des cases assombries en dehors de la diagonale indiquent une plus grande confusion. Une différence négative (valeur d'une case moins

élevée dans la nouvelle matrice) est une case éclaircie et correspond à un affaiblissement de la reconnaissance si la case est sur la diagonale ou à une diminution de la confusion si elle est en dehors.

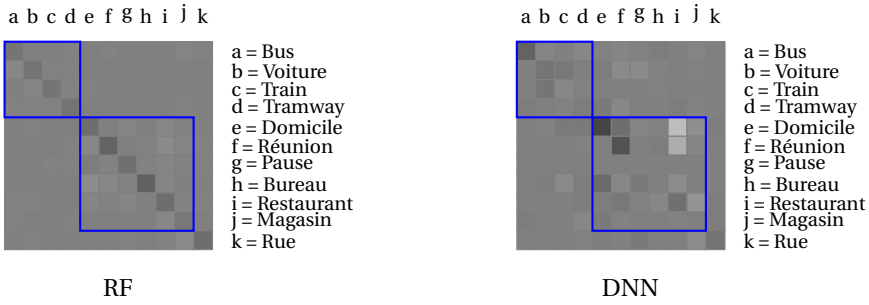


FIGURE 5.7: Représentation des différences, élément par élément entre les matrices de confusio du corpus équilibré *REF* et les matrices du corpus équilibré *REF_SA* pour les classifieurs RF et DNN

Le RF présente une diagonale assombrie et des cases éclaircies en dehors. En particulier, on remarque plusieurs de ces cases dans le carré des environnements en intérieur, ce qui indique une moins grande confusion entre ces environnements. La matrice du DNN est plus composée. La tendance globale des éléments de diagonale indique une augmentation de la reconnaissance, en particulier pour le *bus* (a), le *domicile* (e) et la *réunion* (f). Également, ces deux dernières scènes sont moins confondues avec le *restaurant*. Les variations des autres confusions sont plus hétérogènes.

Nous présentons dans la table 5.8 les confusions recalculées suivant les macro-environnements. L'augmentation du rappel moyen est de 3 points pour les deux classifieurs en comparaison avec les mêmes matrices pour le corpus équilibré *REF* (figure 5.6).

		RF				**DNN**			
Taille		a	b	c	← Reconnu	a	b	c	← Reconnu
Transports	29361	**97,2**	2,3	0,6	a = Transports	**85,0**	14,2	0,8	a = Transports
Intérieur	284036	0,9	**97,6**	1,5	b = Intérieur	10,6	**86,9**	2,5	b = Intérieur
Extérieur	24891	3,9	**9,2**	**86,9**	c = Extérieur	12,3	16,7	**71,0**	c = Extérieur
			93,9		Moyenne		81,0		Moyenne

FIGURE 5.8: Matrices de confusion de la reconnaissance des scènes recalculée pour les groupes de macro-environnements sur le corpus *REF_SA*

5.4.3 Classification sur le corpus *REF_AccAud*

Nous présentons brièvement les résultats de la classification dans le cas de la configuration de capteurs *REF_AccAuc* dont seuls les descripteurs de l'accéléromètre et du microphone sont gardés de la configuration *REF*. Comme cela a déjà été observé dans la table 5.6 (à la fin de la section 5.3.2) qui compare les performances sur les trois corpus en validation croisée stratifiée à 10 sous-ensembles, les performances sont fortement réduites lorsque l'on ne considère que l'accéléromètre et le microphone. Ainsi, le taux de classification est seulement de 71,0 % pour le RF, qui reste le meilleur classifieur. Le DNN chute à 11,9 %, à peine plus que la probabilité d'un tirage aléatoire d'une scène parmi les onze du corpus.

TABLE 5.11: Mesures de performance pour la reconnaissance de scènes à partir de l'accéléromètre et du microphone exclusivement

	C4.5	RF	NB	GMM	DNN
F-mesure moy.	48,2	**58,0**	14,7	23,5	19,9
Rappel moy.	63,9	**73,9**	21,1	34,2	34,7
Précision moy.	47,2	**55,5**	30,4	28,7	32,7
Taux de classification	60,0	**71,0**	10,0	23,5	11,9

La comparaison des nouvelles matrices de confusion avec celles de la configuration *REF*, pour les classifieurs RF et DNN, est illustrée dans la figure 5.9. Elle indique une baisse générale des valeurs sur la diagonale et une augmentation des confusions, en particulier en dehors des macro-environnements. Ainsi, le *train* (ligne *c*) et le tramway (colonne *d*) présentent de nombreuses confusions avec les environnements en intérieur *e* à *j* (plus marquée dans la matrice du DNN).

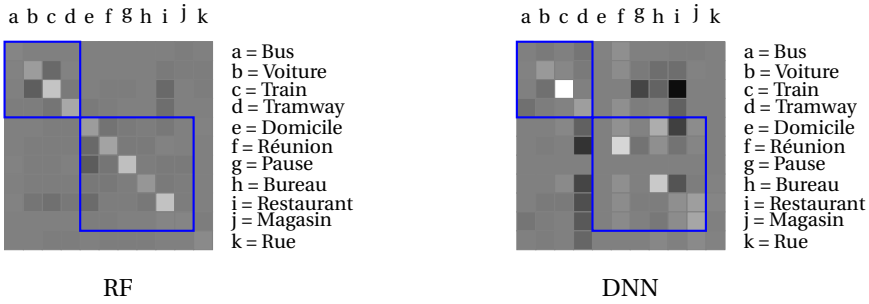


FIGURE 5.9: Différences éléments par éléments des matrices de confusion de reconnaissance de scènes du corpus *REF_AccAud* relativement au corpus *REF* pour les classifieurs RF et DNN

5.5 Détection des transitions

Nous présentons dans cette section les résultats de l'expérimentation de détection des transitions entre scènes. L'expérimentation est justifiée dans la section 4.2.2 par la présence de scènes très courtes et l'hypothèse que les transitions sont identifiables par les changements de lieu et d'action. Notre approche vise à identifier le plus grand nombre de ruptures. Cela risque de créer un nombre important de fausses alarmes que nous supposons pouvoir limiter par la combinaison à un système de classification.

La méthode choisie pour cette expérimentation consiste en un système de segmentation de séquence de vecteurs, initialement créée pour la segmentation en locuteurs de documents sonores. Le fonctionnement théorique de la méthode est introduit dans la dernière section du chapitre de l'état de l'art, aussi nous ne faisons que rappeler les principales étapes de la méthode. La première étape procède de manière itérative sur la séquence de vecteurs. Elle considère deux fenêtres successives de vecteurs afin d'évaluer l'hypothèse la plus vraisemblable de génération de vecteurs par la même distribution ou deux distributions différentes. Une frontière est placée entre deux segments lorsque la seconde hypothèse est la plus

vraisemblable. La seconde étape reproduit le mécanisme de la première mais en considérant les segments obtenus précédemment. De cette manière, on opère une fusion sur certains des segments de la première étape. La troisième phase vise à regrouper les segments similaires non contigus. Pour cela, toute la séquence de vecteurs est considérée.

Nous avons choisi cette méthode car elle est reconnue dans le domaine de la segmentation en locuteurs. Par ailleurs, son fonctionnement est intéressant car il combine une approche générative pour la représentation des segments à une approche discriminante, pertinente pour la détection de ruptures. Cette méthode présente quelques inconvénients comme la nécessité de déterminer empiriquement des paramètres de fonctionnement optimaux. De plus, elle requiert l'ensemble du signal pour affiner la segmentation, ce qui n'est pas réaliste dans le contexte d'une application en continu et en temps-réel (en réalité, la première étape peut être appliquée à un flux en temps-réel car elle fonctionne progressivement sur des fenêtres de vecteurs).

Par ailleurs, nous avons fait le choix de limiter l'expérimentation aux seuls descripteurs de l'accéléromètre composés des valeurs de moyenne et variance sur les 3 axes (résultant en un vecteur de 6 descripteurs). Ce choix est justifié par la nature des transitions, qui représentent des changements de lieu ou d'action. Dans les deux cas, des mouvements sont effectués. L'accéléromètre apparaît comme la source de données la plus opportune pour capturer les variations dans les mouvements. Le gyroscope est un autre bon candidat, mais sa forte consommation d'énergie relativement à l'accéléromètre, sa disponibilité plus incertaine sur les appareils et sa faible présence dans les travaux de l'état de l'art justifient de ne pas l'employer.

L'application de la méthode de segmentation a nécessité la détermination de plusieurs paramètres. Le premier porte sur la taille des fenêtres de vecteurs à considérer dans la première étape de segmentation. Celle-ci doit être suffisante pour permettre l'estimation des paramètres de la distribution. Puisque la valeur représente le nombre de vecteurs consécutifs nécessaire pour une estimation, elle peut être interprétée comme le retard avant d'avoir une prédiction dans l'hypothèse d'un flux continu en temps-réel. Aussi, nous avons souhaité limiter ce retard en vue de l'objectif d'une application industrielle. C'est pourquoi nous avons fixé la taille de la fenêtre à 30 vecteurs (ou 30 secondes entre deux estimations).

Ensuite, pour chacune des deux étapes suivantes de la méthode, il est nécessaire d'estimer le seuil du score de log-vraisemblance pour accepter une fusion. Nous avons choisi de fixer le premier seuil à une valeur jugée optimale. Pour déterminer cette valeur, nous avons effectué plusieurs expérimentations en faisant varier le seuil dans une plage de valeurs et nous avons sélectionné celle pour laquelle le rappel est maximal. Le second seuil est resté libre et a servi de paramètre variable pour évaluer les résultats du système de détection.

Pour l'expérimentation de la détection de transitions, nous disposons du même corpus que précédemment, composé des 22 enregistrements considérés comme indépendants (voir la section 5.1.1 pour plus de détails sur la description du corpus). Les transitions sont marquées par l'étiquette "incertain" et s'étendent sur une période moyenne de 2 minutes. Par ailleurs, le système de détection indique une frontière qui sépare deux segments. Nous

considérons que la frontière du système est correcte si elle se trouve dans une période de transition de la référence. Ainsi, suivant cette considération et la volonté de limiter les vraies transitions manquées, nous évaluons la performance du système par le calcul du score de rappel. Le nombre de transitions au sein desquelles le système a indiqué une ou plusieurs frontières représente le nombre de transitions correctement identifiées. De plus, nous calculons le *taux de segmentation* de la proposition du système sur tout l'enregistrement par le nombre de frontières indiquées par le système sur le nombre de transitions à découvrir. Un taux supérieur à 1 indique une sur-segmentation tandis qu'un taux inférieur indique une sous-segmentation.

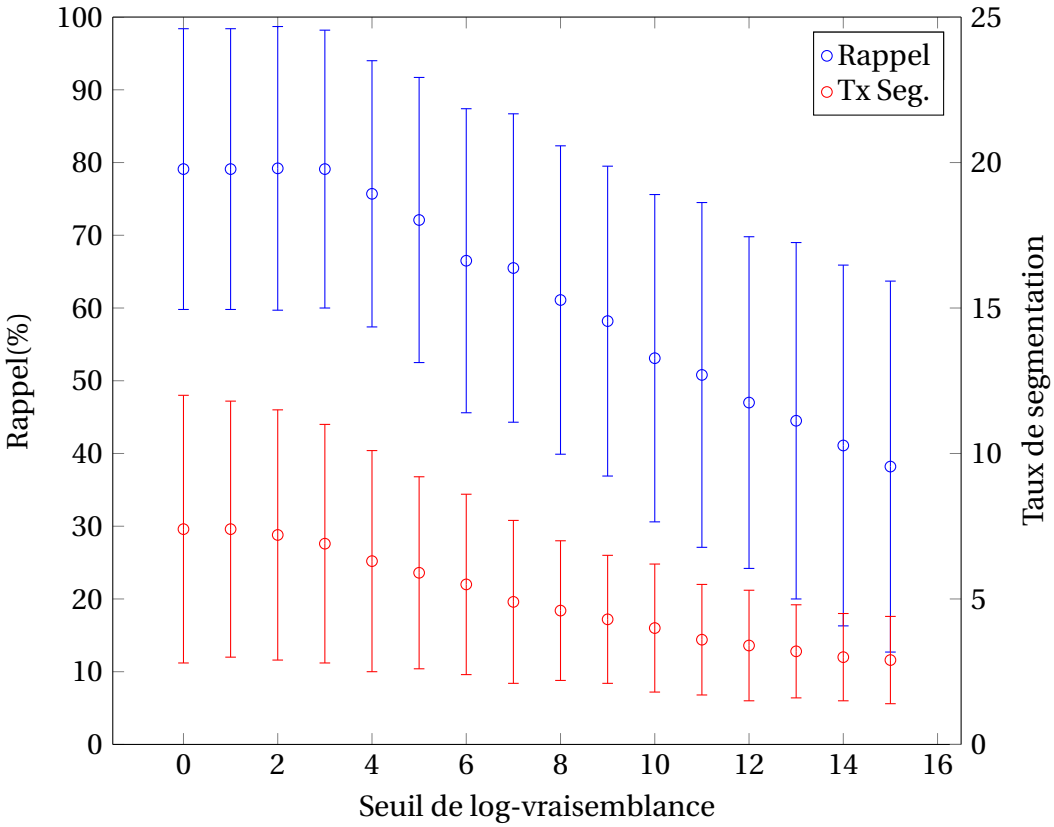


FIGURE 5.10: Rappel et taux de segmentation suivant le seuil de log-vraisemblance de la méthode de segmentation

Pour l'expérimentation, le système est appliqué sur chaque enregistrement en faisant varier le second seuil. Le rappel moyen sur tous les enregistrements et l'écart-type des valeurs sont calculés et représentés dans la figure 5.10. Les valeurs sont interprétées suivant l'axe de gauche du graphique. Le taux de segmentation moyen et l'écart-type des valeurs est également représenté sur le même graphique, dont les valeurs s'échelonnent suivant l'axe de droite. La tendance observée sur le graphique est prévisible. L'augmentation du seuil d'acceptation de fusion de segments correspond à une tolérance plus forte des scores pour la fusion et augmente ainsi le nombre de fusions effectuées. En augmentant les fusions, le nombre de segments restants diminue et le taux de segmentation avec. Par ailleurs, le risque augmente de fusionner des segments dont la frontière représente effectivement une transition, ce qui a tendance à diminuer le score de rappel. Les scores de rappel les plus élevés

atteignent presque 79 % en moyenne avec un écart-type qui ne descend pas sous la barre des 60 %, ce qui représente des valeurs très encourageantes. Le taux de segmentation correspondant est proche de 7, ce qui indique 7 fois plus de segments identifiés que dans la référence de l'annotation, mais comme nous l'avons évoqué, nous faisons l'hypothèse qu'un classifieur peut le réduire en "lissant" la proposition du système de segmentation.

5.6 Bilan

Nous concluons sur ce chapitre en commençant par les résultats de la sélection d'attributs. Nous avons mis en évidence que la moyenne de la pression atmosphérique est le descripteur le plus pertinent de l'ensemble. Nous retenons aussi la pertinence des sources inertielles et du microphone, dont plusieurs descripteurs sont retenus.

Nous présentons dans les figures 5.11 et 5.12 les valeurs de rappel suivant les configurations de capteurs et dans les deux conditions d'évaluation. Les classifieurs à base d'arbres obtiennent les meilleurs résultats et la forêt d'arbres décisionnels RF est peu sensible à la réduction de descripteurs. Les moins bons résultats des GMM et DNN peuvent s'expliquer par le manque de données dans le corpus ainsi que par la faible adaptation des paramètres des modèles. Enfin, le résultat mauvais et constant du NB conforte l'idée que l'hypothèse sous-jacente de dépendance directe conditionnelle de la scène aux descripteurs est invalide.

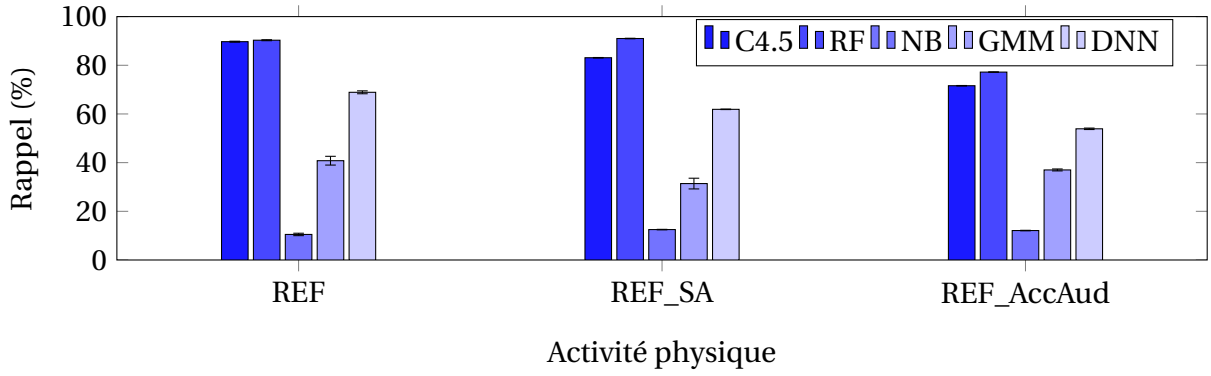


FIGURE 5.11: Rappel de classification en validation croisée stratifiée à 10 sous-ensembles

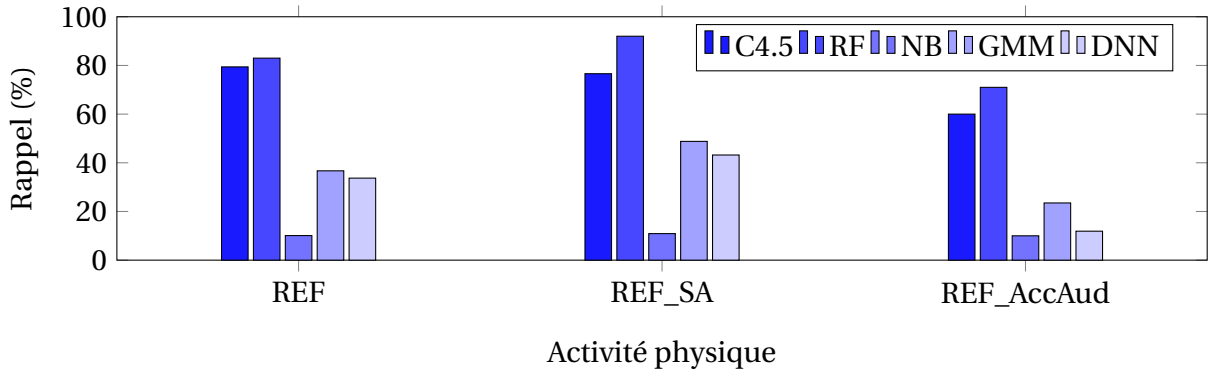


FIGURE 5.12: Rappel de classification après entraînement sur le corpus uniforme

Nous avons aussi vu dans le chapitre que les résultats peuvent s'améliorer lorsque l'on considère les macro-environnements. Ce résultat est intéressant pour l'application industrielle. En effet, il s'agit d'abord d'une description de scène complémentaire à l'étiquette de scène, ce qui est en accord avec la volonté industrielle d'obtenir plusieurs niveaux de description. En outre, cela permet d'envisager l'usage du service par l'application d'un compromis entre performance et précision de la description.

Enfin, nous avons présenté une expérimentation de détection de transitions qui a donné une mesure de rappel de 80 % et une très forte sur-segmentation du signal. Le système de détection n'a pas été évalué en combinaison avec un classifieur lisseur, mais nous suggérons que les résultats pourraient s'améliorer. Le bilan termine la première partie des expérimentations. Les résultats présentés représentent une référence, obtenus dans des conditions d'exploitation de capteurs et de corpus précises, avec un ensemble de classifieurs de l'état de l'art. Les résultats sont encourageants, aussi dans le chapitre suivant, nous proposons des expérimentations complémentaires qui visent à compléter la compréhension du modèle de scène et à proposer un système de reconnaissance alternatif.

CHAPITRE 6

Analyse exploratoire des données de scènes

Les deux problèmes principaux de la thèse ont été abordés dans les chapitres précédents. D'abord, la définition du concept de scène a été traitée par la confrontation de travaux existants à une analyse qualitative des annotations collectées. Concernant le problème de la reconnaissance de scène, le chapitre précédent a présenté une approche commune s'appuyant sur l'apprentissage supervisé et évalué dans plusieurs configurations pertinentes.

Dans ce chapitre, nous proposons une approche alternative pour ces deux problèmes. D'abord, nous abordons la question de la définition d'une scène par l'étude de concepts naturels dans les données de scènes. L'émergence de ces concepts est effectuée de manière non-supervisée et l'interprétation vise à identifier les concepts obtenus pour mieux expliquer la notion de scène ou certaines scènes particulières.

Ensuite, nous proposons un système de reconnaissance alternatif, qui repose sur la notion de composition d'une scène. Le système est constitué de deux classifieurs, dédiés à la reconnaissance du lieu et de l'action (les deux éléments identifiés qui composent une scène suivant la définition du chapitre 4). L'estimation de la scène est effectuée par la fusion des prédictions des classifieurs. La pertinence de ce système réside *a priori* dans la description de la scène fournie en sortie, plus complète que dans l'approche de classification du chapitre précédent, et donc plus en accord avec le cahier des charges.

6.1 Analyse non-supervisée des données

Les expérimentations présentées dans cette section visent deux objectifs. Le premier porte sur l'interprétation des données de situations collectées afin de faire émerger des concepts pour expliquer les scènes. Le deuxième objectif consiste à quantifier les interprétations par des valeurs numériques (par exemple, des probabilités) afin de les utiliser dans un système plus complexe de reconnaissance de scènes.

6.1.1 Analyse des groupes de vecteurs d'accélération

Pour les trois expérimentations d'interprétation, nous considérons le même corpus composé de données d'accélération et d'ambiance sonore. Nous reprenons le corpus présenté dans la section 5.1 enrichi d'un enregistrement supplémentaire. Le corpus total représente 434792 secondes, dont la répartition est décrite dans la table 6.1.

TABLE 6.1: Détails du corpus employés pour l'expérimentation d'exploration des données

Scène	**Instances**	**Durée (s)**	**Durée (h)**
Bateau	1	1548	0,4
Train	1	4200	1,2
Pause	8	10076	2,8
Magasin	12	12085	3,4
Voiture	16	32239	9,0
Tram	27	16571	4,6
Bus	11	15225	4,2
Rue	96	33495	9,3
Réunion	11	52956	14,7
Restaurant	15	46400	12,9
Domicile	30	92911	25,8
Bureau	21	117086	32,5
Total	249	434792	120,8
Incertitude	*211*	*30569*	*6,5*

Pour l'interprétation des groupes à partir des seules données d'accélération, l'expérimentation consiste à regrouper les vecteurs de manière non-supervisée en appliquant une méthode de regroupement qui repose sur l'algorithme EM. La méthode est décrite dans le chapitre 2, à la section 2.4.3.2. Brièvement, la méthode estime les paramètres d'un GMM dont les gaussiennes représentent les groupes et détermine le nombre optimal de groupes de manière itérative. En commençant avec un groupe unique (soit une seule gaussienne), le corpus est découpé en 10 sous-ensembles. À la manière d'une validation croisée, 9 sous-ensembles sont employés pour appliquer l'algorithme EM afin d'estimer les paramètres de la gaussienne. Puis on calcule la log-vraisemblance que la gaussienne ait généré les vecteurs du dixième sous-ensemble. On répète l'opération 9 fois, de sorte que chaque groupe serve au calcul de la log-vraisemblance. À la fin, la log-vraisemblance moyenne est calculée. Si la différence de log-vraisemblance avec l'étape précédente est positive, alors on incrémente d'une unité le nombre de groupes et on recommence la phase de validation croisée sur le nouveau nombre de groupes. La méthode s'arrête lorsque la log-vraisemblance n'augmente pas. On garde les groupes de la dernière itération.

Les vecteurs considérés dans cette première expérimentation sont les moyennes et variances de la composante d'accélération projetée sur les trois axes du repère. Soit W la fenêtre de longueur n appliquée sur la séquence continue des échantillons d'un enregistrement. Les échantillons sélectionnés par la fenêtre sont des mesures d'accélération $\vec{a}_i$, chacune représentée sous forme de vecteur contenant la projection de l'accélération suivant les trois axes du repère. Nous assimilons la fenêtre W à la séquence d'échantillons sélectionnés : $W = [\vec{a}_1, \vec{a}_2, \ldots, \vec{a}_n]$. La moyenne et la variance sont calculées sur cette fenêtre suivant les deux équations :

$$\begin{aligned} \vec{\mu_n} &= \frac{1}{N}\sum_{i=1}^{n}\vec{a}_i \\ \vec{\sigma_n}^2 &= \frac{1}{N}\sum_{i=1}^{n}(\vec{a}_i - \vec{\mu_n})^2 \end{aligned} \tag{6.1}$$

Dans la suite, un vecteur de descripteurs d'accélérations est la concaténation de ces deux vecteurs de paramètres et contient six descripteurs au total.

Afin de pouvoir interpréter les groupes de vecteurs, nous introduisons dans la figure 6.1 le repère orthonormé considéré pour les mesures d'accélération. Nous précisons également que les mesures d'accélération tiennent compte de l'accélération de pesanteur (ou gravité) suivant le principe fondamental de la dynamique du point, exprimé par la formule suivante :

$$\vec{a_{\text{acc}}} = -\vec{g} - \frac{1}{m}\sum \vec{f} \tag{6.2}$$

où $\vec{a_{\text{acc}}}$ représente l'accélération mesurée par l'accéléromètre, $\vec{g}$ l'accélération de pesanteur terrestre [1], $\sum \vec{f}$ la somme des forces appliquées au capteur et m la masse de celui-ci. Ainsi, lorsque le téléphone est au repos, aucune force ne s'applique sur l'accéléromètre et l'accélération mesurée provient de la seule force de gravité. Si le téléphone est posé sur une surface plane et horizontale, l'écran vers le haut, alors l'accélération est projetée sur l'axe z et sa valeur vaut $a_{\text{acc}} = -g = 9,81 m/s^2$. L'orientation des vecteurs est relative au repère de la figure 6.1.

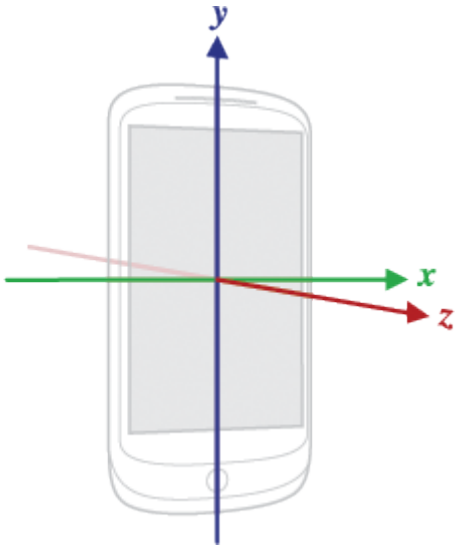


FIGURE 6.1: Système de coordonnées employé pour la mesure d'accélération sur un smartphone

L'application de la méthode sur les vecteurs d'accélérations a permis d'obtenir trois groupes. Dans la table 6.2, nous décrivons les projections des centroïdes sur chacun des six descripteurs ainsi que l'écart-type des valeurs à la projection. Par exemple, la projection du centroïde du groupe 0 sur le descripteur de moyenne de l'accélération de l'axe x est centré sur la valeur $-0,1235 m/s^2$ avec un écart-type des valeurs de $\pm 0,2151 m/s^2$.

La première observation porte sur le groupe 0 dont le centroïde affiche une valeur de 9,7266 m/s^2 pour le descripteur de moyenne d'accélération projetée sur l'axe z. Cette valeur est très proche de l'accélération de gravité . En outre, les valeurs des moyennes d'ac-

1. estimée dans le repère du téléphone à $-9,81 m/s^2 \vec{z}$ d'après la documentation des développeurs Android https://developer.android.com/guide/topics/sensors/sensors_motion.html#sensors-motion-accel

TABLE 6.2: Projection des centroïdes des groupes de vecteurs suivant sur les six descripteurs d'accélération

	Nb. vect.	moy. axe x	moy. axe y	moy. axe z	var. axe x	var. axe y	var. axe z
gpe. 0	174782	-0,1235 ± 0,2151	0,0909 ± 0,2719	**9,7266** ± 0,0516	**0,0002 ± 0,0001**	**0,0002 ± 0,0001**	**0,0003 ± 0,0002**
gpe. 1	136813	-0,3273 ± 2,5111	-1,7271 ± 4,6846	1,0150 ± 8,1867	0,0035 ± 0,0063	0,0019 ± 0,0033	0,0036 ± 0,0068
gpe. 2	153766	0,3766 ± 3,5421	-2,5605 ± 6,6980	2,5275 ± 4,8126	**1,4954 ± 2,7895**	**2,4905 ± 4,8111**	**2,1998 ± 4,4221**

célérations sur les deux autres axes sont proches de 0 et les variances d'accélérations sont presque nulles. Pour interpréter ces observations, nous faisons l'hypothèse que le groupe représente des moments d'immobilité où le smartphone est posé sur une surface horizontale avec l'écran tourné vers le haut.

La répartition de variance d'accélération du groupe 1 est très similaire à celle du groupe 0 : les valeurs sont très faibles et les écarts-types de ces valeurs sont également faibles. Cependant, le groupe 1 se distingue par les descripteurs de la moyenne d'accélération, dont les écarts-types sont très élevés. Pour expliquer ces deux observations, nous suggérons que les vecteurs rassemblés dans ces groupes représentent des moments d'immobilité de l'appareil (voire de l'utilisateur) au cours desquels l'appareil est maintenu dans une orientation autre que l'horizontalité. En effet, pendant l'immobilité, seule la gravité terrestre est mesurée. Par conséquent, une telle orientation mène à des projections de la gravité sur les 3 axes. En outre, pour expliquer les grands écarts-types observés, nous suggérons que le groupe 1 rassemble plusieurs sous-groupes d'immobilité aux orientations différentes et donc aux différentes projections de la gravité sur les 3 axes. Des exemples de tels sous-groupes sont la tenue du téléphone à l'oreille pendant un appel, qui correspond à la fois à un faible mouvement et à une orientation particulière ; ou encore le rangement de l'appareil dans un sac pendant une réunion.

Enfin, le groupe 2 se distingue des deux groupes précédents par les valeurs élevées de variance d'accélération sur les 3 axes. Par définition, cela indique que la mesure de l'accélération sur un axe a varié relativement à la moyenne. Cette variation peut être causée soit par un changement d'orientation rapide du téléphone soit par l'application d'une force sur l'appareil. Le premier cas peut être illustré par l'exemple de la marche, avec le téléphone dans la poche du pantalon. Avec le mouvement périodique de la jambe, l'orientation du téléphone varie régulièrement, ce qui fait évoluer la projection de la gravité sur les axes et donc la mesure de l'accélération. Le deuxième cas est illustré par un déplacement en transport (une voiture par exemple). Le téléphone est posé ou rangé, mais le déplacement du véhicule crée des forces (telle que la force centrifuge dans les virages) qui peuvent être perçues par le téléphone. En outre, suivant l'état de la voiture, celui de la route et la position du téléphone, des vibrations peuvent être mesurées et créer des variations dans les mesures d'accélération. Les valeurs de moyenne et de variance d'accélération du groupe 2 ne nous permettent pas de préciser nos hypothèses.

Pour résumer et faciliter la représentation dans la suite, nous associons un label à chaque groupe :

— le groupe 0 est associé au label "posé" qui indique que le téléphone est posé à plat sur

une surface plane ;

— le groupe 1 est associé à l'étiquette "calme" qui correspond à l'immobilité du téléphone ou à de très faibles mouvements, avec une orientation différente de celle du groupe 0 ;

— le groupe 2 est associé au label "agité" qui rappelle l'importante variance d'accélération observée.

Par la suite, dans la table 6.3, nous décrivons la répartition des vecteurs de scène dans chaque groupe. Les valeurs de la table se lisent dans le sens horizontal.

TABLE 6.3: Répartition des scènes dans les groupes

	Nb. vect.	Bateau	Bus	Voiture	Domicile	Réunion	Pause	Bureau	Restau.	Magasin	Rue	Train	Tramway	Incertain
posé	174782	0.0	0.0	0.0	**28.9**	**15.6**	1.0	**49.1**	4.5	0.2	0.0	0.3	0.1	0.4
calme	136813	0.9	4.4	1.4	**21.4**	**13.5**	3.5	**16.1**	**16.6**	4.2	3.8	1.3	8.0	4.9
agité	153766	0.2	6.0	**19.7**	8.5	4.7	2.4	5.9	**10.3**	3.9	**18.4**	1.2	3.6	**15.2**

Il apparaît que le groupe *posé* est très présent dans la scène du *bureau* (49,1 % des vecteurs du groupe pour cette scène seule), du *domicile* avec 28,9 % et de la *réunion* (15,6 %). Dans une moindre mesure, la scène du *restaurant* contient 4,5 % des données du groupe. Suivant l'hypothèse faite sur ce groupe, ces scènes apparaissent favorables à la pose du téléphone sur une surface plane tandis que les autres scènes ne le sont pas.

Les vecteurs du groupe *calme* se répartissent aussi dans le *domicile*, le *restaurant*, le *bureau* et la *réunion*. Les quatre scènes regroupent 67,6 % des vecteurs. La représentation intuitive de ces scènes correspond à la présence de faibles mouvements. Par exemple, les personnes peuvent porter le smartphone dans une poche, dans une posture assise.

Les vecteurs du groupe *agité* se répartissent dans les scènes de *voiture*, *rue*, *restaurant* et les transitions entre scènes qui portent l'étiquette "incertain". Ces quatre scènes cumulent 63,6 % des vecteurs du groupe.

Nous remarquons que la scène de la *rue*, du *bureau*, du *domicile*, de la *réunion* et du *restaurant* sont les plus représentées par les vecteurs de descripteurs et cumulent 78,9 % du total des vecteurs. Ce taux éclaire la faible répartition des autres scènes dans les groupes.

Afin de compléter l'interprétation des scènes par les groupes, nous proposons dans la table 6.4 la répartition des groupes dans chacune des scènes. Contrairement à la table précédente, le sens de lecture est vertical.

TABLE 6.4: Répartition des groupes pour chaque scène

	Domicile	Réunion	Bureau	Restau.	Pause	Train	Magasin	Bateau	Tramway	Bus	Incert.	Rue	Voiture
posé	54,4	51,4	73,4	16,8	16,7	12,8	2,9	0,0	0,7	0,0	2,0	0,0	0,0
calme	31,5	34,9	18,8	49,0	47,2	43,6	47,3	79,4	65,7	39,3	21,7	15,6	6,1
agité	14,0	13,7	7,8	34,2	36,1	43,6	49,8	20,6	33,5	60,7	76,3	84,4	93,9
Nb. vect.	92911	52956	117086	46400	10076	4200	12085	1548	16571	15225	30569	33495	32239

L'observation générale de la table montre l'absence d'association forte entre un groupe et une scène. À l'inverse, toutes les scènes sont composées de deux ou trois groupes. Nous avons organisé les compositions de scènes en quatre ensembles, marqués par l'alternance

de couleur de fond grisée ou blanche. Nous commençons par les scènes du *bureau*, de la *réunion* et du *domicile*. Ces scènes sont composées pour majorité des vecteurs des groupes *posé* et *calme*, ce qui est en accord avec les observations précédentes. En outre, nous remarquons la présence non négligeable des vecteurs du groupe *agité*. Cette observation n'est pas surprenante car ces scènes peuvent donner lieu à de l'agitation tels que des déplacements.

Les scènes du *bus*, de la *rue*, de la *voiture* ainsi que les transitions entre scènes sont majoritairement composées des vecteurs du groupe *agité*. Les vecteurs du groupe *calme* complètent la composition. Ces scènes sont aussi caractérisées par un nombre très faible voire nul de vecteurs du groupe *posé*. Cela signifie que le téléphone n'y est pas posé sur une surface plane, immobile. Si l'on suppose que les situations de la *rue* qui ont été collectées correspondent principalement à de la marche, alors on peut suggérer que le point commun de ces groupes est le déplacement. Les vecteurs du groupe *calme* peuvent représenter des attitudes immobiles assise ou debout, dans les transports comme le *bus* ou la *voiture* ou dans la rue, à l'attente à un croisement, par exemple. Les vecteurs du groupe *agité* et des transitions entre scènes peuvent représenter des situations de marche, des changements de trajectoires dans les transports ou des vibrations du véhicule ou de la route.

Le *bateau* et le *tramway* sont également des scènes de déplacement et leur composition n'est pas en contradiction avec celles des quatre scènes précédentes. La différence notable réside dans la proportion de vecteurs des groupes *calme* et *agité*, qui est inversée : le groupe *calme* est prédominant dans ces scènes. Si l'on considère d'une part que la situation du *bateau* représentée dans les données est un ferry (donc un bateau grand et large, moins sensible aux mouvements des vagues) ; et d'autre part que le *tramway* suit des trajectoires plus linéaires que d'autres moyens de transport comme le *bus* ou la *voiture*, alors la prédominance du groupe *calme* sur celle du groupe *agité* paraît vraisemblable.

La même interprétation peut être faite pour les vecteurs du *train*, ce qui permet d'expliquer la prédominance des groupes *calme* et *agité*. Les vecteurs du groupe *posé* dans cette scène peuvent être expliqués par la pose du téléphone sur la tablette d'un wagon lorsque le train s'arrête dans une gare. La présence de vecteurs du groupe *posé* dans les scènes du *restaurant* et de la *pause* est justifiée par la possibilité de poser le téléphone dans ces scènes. Les vecteurs du groupe *agité* dans ces scènes et dans celle du *magasin* peuvent correspondre à des déplacements à pied, potentiellement fréquents. Enfin, les vecteurs du groupe *calme* correspondent aux attitudes immobiles.

Afin de vérifier la cohérence de ces trois groupes, nous avons testé une classification supervisée. La table 6.5 présente la matrice de confusion de la classification. L'expérimentation est effectuée en validation croisée à 10 sous-ensembles et la mesure de rappel est moyennée sur les 10 répétitions. Nous avons choisi l'arbre de décision C4.5 pour la classification, entraîné avec élagage et avec un minimum de 100 vecteurs par feuille avec l'outil Weka. Ce classifieur est intéressant car sa stratégie d'apprentissage diffère de celle de l'algorithme de regroupement employé. Les valeurs très élevées de la diagonale montrent la cohérence des trois groupes et la capacité de l'algorithme à les distinguer.

Nous présentons aussi l'arbre de décision dans la figure 6.2. Les couleurs des feuilles re-

TABLE 6.5: Matrice de confusion de la classification des 3 groupes de vecteurs d'accélération

posé	calme	agité	← classé comme
99,8	0,2	0	posé
0,7	98,8	0,5	calme
0	0,4	99,6	agité

présentent les groupes associés : gris pour le groupe *posé*, jaune pour le groupe *calme* et orange pour le groupe *agité*. Tous les vecteurs associés à une feuille n'appartiennent pas nécessairement au même groupe. C'est pourquoi les feuilles indiquent deux valeurs : le nombre de vecteurs du groupe majoritaire puis le nombre de vecteurs des autres groupes. Nous avons représenté par un contour bleu les attributs de moyenne d'accélération. La moyenne en z apparaît comme un attribut pertinent car on la retrouve sur la deuxième couche de nœuds de l'arbre. Cependant, la majorité des tests impliquent les descripteurs de variance, qui sont donc les plus discriminants.

Par ailleurs, on constate que les feuilles les plus importantes associées au groupe *agité* (orange) sont situées dans les premières couches de l'arbre, ce qui indique une caractérisation aisée du groupe. Le groupe *posé* (gris) requiert de nombreux tests mais le nombre de feuilles associées est limité, ce qui indique une cohérence forte des vecteurs qui le composent. Enfin, les feuilles du groupe *calme* sont les plus nombreuses et nécessitent de nombreux tests pour être différenciées des feuilles des deux autres groupes. Cette observation est cohérente avec l'hypothèse émise sur la composition de plusieurs sous-groupes aux orientations différentes.

Pour conclure sur cette première expérimentation, nous avons mis en évidence trois groupes de données d'accélération, issus d'un regroupement non-supervisé. Par l'étude des centroïdes des groupes, nous avons émis des hypothèses sur l'orientation et la quantité de mouvement associées aux groupes. Puis l'observation de la répartition dans les scènes a permis d'émettre des hypothèses sur les actions associées aux groupes. L'expérimentation de classification a confirmé la cohérence des groupes et la capacité à les différencier. Dans la suite, nous souhaitons effectuer la même analyse avec les groupes issus du regroupement de vecteurs acoustiques.

6.1.2 Analyse des groupes de vecteurs acoustiques

Les vecteurs acoustiques sont composés de 40 coefficients d'énergie issus de filtres répartis de manière linéaire dans l'échelle logarithmique des Mel, dans la bande de fréquences équivalente de 0 à 22050 Hz, . Comme pour les descripteurs d'accélération, les coefficients sont calculés sur les échantillons des fenêtres de signal. Les fenêtres sont plus courtes (1024 échantillons, soit 23 millisecondes environ) et les coefficients sont moyennés sur plusieurs fenêtres consécutives afin de couvrir les fenêtres de 2 secondes synchronisées avec celles des descripteurs d'accélérations.

L'application de la méthode sur les vecteurs acoustiques a permis d'obtenir neuf groupes. Nous présentons dans la table 6.6 leur répartition dans chaque scène. L'observation géné-

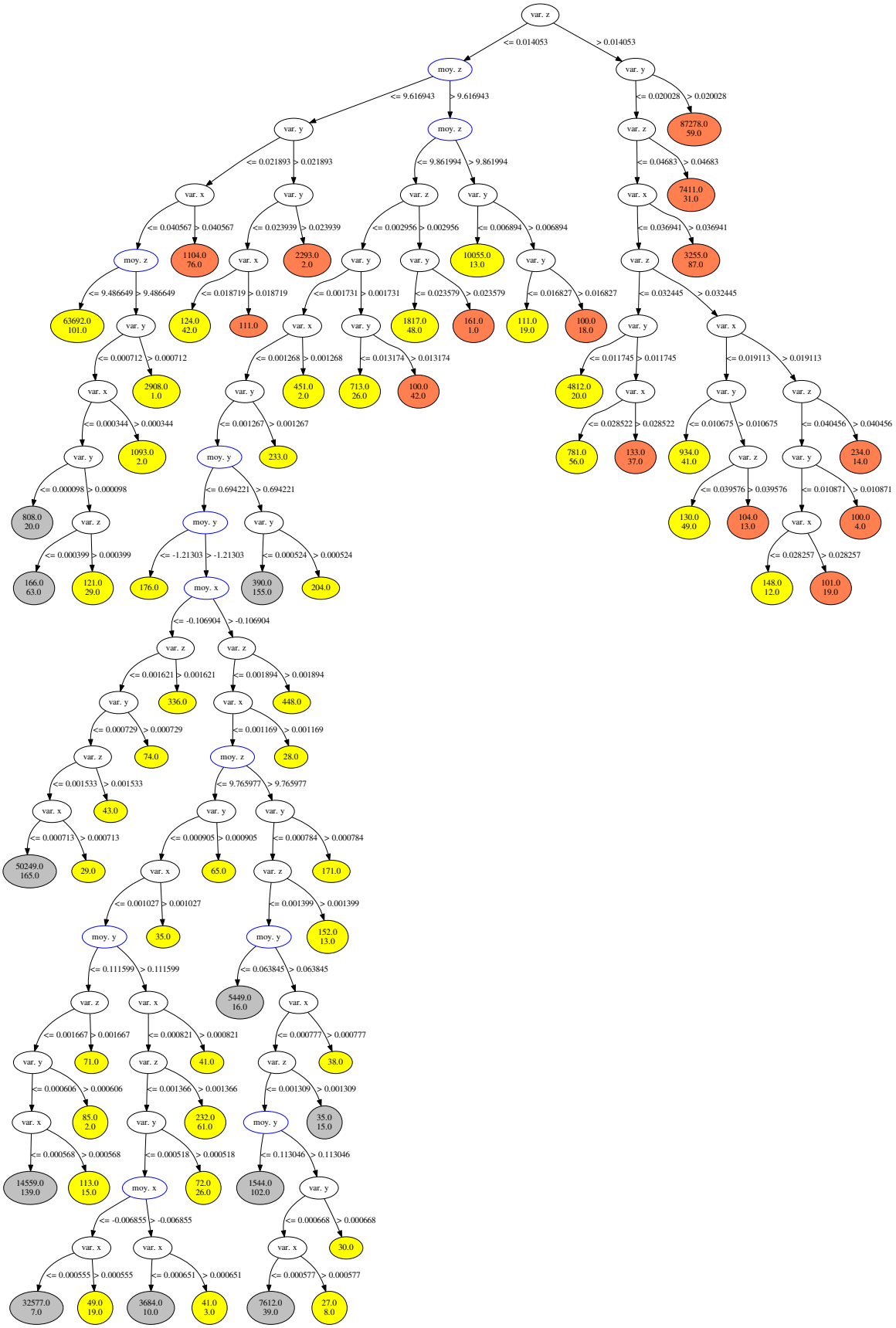


FIGURE 6.2: Arbre de décision C4.5 entraîné à reconnaître les trois groupes *posé* (en gris), *calme* (en jaune) et *agité* (en orange)

rale indique l'absence d'association forte entre les scènes et les groupes, à l'exception de la scène de *train*, presque entièrement décrite par le groupe 6. Les groupes 4 et 6 sont prédominants et contiennent 282736 vecteurs soit 60 % de l'ensemble. Ils sont présents dans de nombreuses scènes, aux ambiances acoustiques *a priori* différentes telles que la *voiture*, le *domicile*, le *bureau*, la *rue* ou le *magasin*.

TABLE 6.6: Répartition des groupes dans les scènes

	Nb. vect.	Bateau	Bus	Voiture	Domicile	Réunion	Pause	Bureau	Restau.	Magasin	Rue	Train	Tramway	Incertain
groupe 0	41308	0,6	**29,6**	5,4	5,2	6,9	7,0	3,1	6,1	**37,4**	20,7	0,3	**23,0**	13,6
groupe 1	6599	0,0	1,8	0,0	0,8	0,2	0,4	0,2	5,1	2,0	4,8	0,0	0,6	2,8
groupe 2	1835	0,1	0,4	0,0	0,3	0,2	1,8	0,1	0,7	0,5	1,1	0,0	0,5	0,6
groupe 3	20981	0,3	7,1	0,3	3,0	1,4	4,6	1,5	9,3	8,5	16,1	0,0	3,2	8,8
groupe 4	125396	24,2	13,8	**61,8**	28,4	14,4	6,6	**50,7**	0,0	**21,8**	2,3	0,2	6,9	14,2
groupe 5	25625	**58,7**	3,9	2,4	3,6	9,4	**20,6**	4,6	0,4	1,5	5,3	0,0	8,0	13,3
groupe 6	157340	0,0	0,0	**29,1**	**39,5**	**32,9**	18,0	22,7	**76,5**	8,3	**32,1**	**98,5**	**30,7**	**29,4**
groupe 7	9044	2,6	1,0	0,1	1,1	1,6	14,4	1,2	0,7	0,8	7,7	0,0	2,0	2,5
groupe 8	77233	13,4	**42,4**	0,9	18,0	**33,0**	**26,6**	15,8	1,2	19,2	9,9	0,9	**25,2**	14,7
Nb. vect.	465361	1548	15225	32239	92911	52956	10076	117086	46400	12085	33495	4200	16571	30569

Les annotations collectées ne permettent pas de décrire précisément l'ambiance sonore des scènes. Nous n'avons pu effectuer que quelques analyses, basées sur des hypothèses simples. Par exemple, nous avons fait l'hypothèse que les scènes de transport peuvent présenter plus d'énergie dans les coefficients de bandes de fréquences basses à cause des bruits de moteurs. Pour cela, nous avons étudié des histogrammes des coefficients d'énergie dans des groupes de vecteurs de scènes de transports. Nous avons observé des amplitudes de ces coefficients particulièrement élevées dans ces coefficients et beaucoup moins dans les coefficients suivants. L'observation a aussi été effectuée dans plusieurs autres scènes de transport. Cependant, puisque nous n'avons pas d'annotations plus précises, nous ne pouvons pas confirmer ces hypothèses ou les approfondir. À titre d'exemple, nous présentons dans la figure 7.1, en annexe, les histogrammes des deux premiers coefficients d'énergie dans les vecteurs du groupe 4 de la scène de voiture.

Malgré le manque d'annotations, nous présentons dans la table 6.7 la matrice de confusion de l'évaluation d'un arbre de confusion sur le corpus annoté avec les étiquettes des groupes. L'expérimentation a été effectuée en validation croisée à 10 sous-ensembles. On remarque des taux de classification très élevés pour tous les groupes, à l'exception des groupes 1 et 2 (qui représentent moins de 2 % des données). Malgré la difficulté d'interprétation des groupes, les valeurs élevées de la matrice indiquent une cohérence dans les groupes obtenus de façon non-supervisée.

6.1.3 Analyse des groupes de vecteurs d'accélérations et acoustiques

Pour compléter l'étude du regroupement non-supervisé, nous décrivons les groupes obtenus lorsque les descripteurs acoustiques et d'accélérations sont réunis. Comme précédemment, nous présentons dans la table 6.8 la répartition des groupes dans chaque scène.

Comme dans les sections précédentes, on remarque l'absence d'association marquée de scènes et de groupes. Toutes les scènes sont composées d'un minimum de 2 groupes. On

TABLE 6.7: Matrice de confusion de la classification des groupes de vecteurs acoustiques

g 0	g 1	g 2	g 3	g 4	g 5	g 6	g 7	g 8	← classé comme
88.9	0.0	0.0	3.4	0.0	2.8	0.0	0.2	4.8	g 0
0.3	**77.4**	3.3	**12.0**	0.0	0.1	0.0	6.9	0.1	g 1
1.1	**22.1**	**64.4**	2.0	0.0	0.3	0.0	**9.5**	0.5	g 2
9.0	1.8	0.0	**84.5**	0.0	1.8	0.0	2.6	0.2	g 3
0.0	0.0	0.0	0.0	**98.0**	0.1	0.7	0.0	1.3	g 4
4.6	0.0	0.0	1.2	1.1	**86.1**	0.0	0.9	6.2	g 5
0.0	0.0	0.0	0.0	0.2	0.0	**99.8**	0.0	0.0	g 6
2.6	2.5	1.0	**8.8**	0.0	5.4	0.0	**79.3**	0.3	g 7
2.1	0.0	0.0	0.0	2.6	1.2	0.0	0.0	**94.1**	g 8

TABLE 6.8: Distribution des groupes de vecteurs acoustiques et d'accélération dans chaque scène

	Nb. inst.	Bateau	Bus	Voiture	Domicile	Réunion	Pause	Bureau	Restau.	Magasin	Rue	Train	Tramway	Incertain
gpe. 0	30831	0.0	0.0	9.3	2.5	3.4	2.0	1.0	**14.4**	0.4	**25.0**	**18.7**	1.8	**18.7**
gpe. 4	10554	1.0	2.4	0.1	1.4	0.8	8.9	0.5	6.6	2.5	7.2	0.0	1.5	3.2
gpe. 7	27450	2.9	2.5	0.4	2.7	2.5	**13.0**	2.8	1.4	9.2	**34.8**	0.1	2.4	**15.1**
gpe. 1	46378	**71.3**	**25.8**	**64.0**	3.9	3.1	5.6	2.8	0.3	9.5	7.3	1.0	6.4	**22.2**
gpe. 6	37917	0.6	**30.3**	5.4	5.9	6.0	**12.6**	2.8	**12.5**	**32.1**	8.5	0.3	**15.5**	**10.8**
gpe. 2	68684	0.0	5.5	0.0	**20.5**	**20.5**	**11.1**	**29.9**	0.0	4.3	1.3	0.0	0.5	2.7
gpe. 3	46083	0.0	0.0	0.0	**16.9**	3.9	0.4	**24.0**	0.0	0.0	0.0	0.0	0.0	0.4
gpe. 5	69589	**24.2**	**33.5**	0.5	**12.5**	**30.2**	**30.4**	12.8	2.7	**20.1**	8.7	0.2	**43.1**	**15.0**
gpe. 8	128325	0.0	0.0	**20.3**	**33.8**	**29.5**	**15.9**	**23.5**	**62.0**	**21.9**	7.2	**79.7**	**28.9**	**11.8**
Taille	465361	1548	15225	32239	92911	52956	10076	117086	46400	12085	33495	4200	16571	30569

remarque néanmoins que les scènes du *domicile*, de la *réunion*, du *bureau* et de la *pause* se répartissent majoritairement dans les groupes 2, 3, 5 et 8. Le groupe 1 est très présent dans plusieurs scènes de transports (*bateau*, *bus* et *voiture*) ainsi que dans les transitions entre scènes. Le groupe 8 est commun à de nombreuses scènes.

Dans la table, nous avons marqué des séparations qui créent 3 ensembles de groupes. Les séparations sont issues de l'étude des descripteurs d'accélération dans chacun des groupes. À la manière de la section 6.1.1, nous proposons une interprétation des centroïdes des groupes suivant les descripteurs de moyenne et de variance d'accélération sur les 3 axes. La figure 6.3 présente la projection des centroïdes de chaque groupe dans les 6 descripteurs d'accélérations, sous forme de diagramme à barres. La barre indique la projection du centroïde pour le descripteur considéré et le segment qui l'encadre indique l'écart-type des valeurs.

L'observation des variances d'accélération permet de distinguer 3 tendances. Les groupes 0, 4 et 7 présentent les valeurs de variance les plus élevées ainsi que les plus grands écarts-types autour de ces valeurs, indiquant une grande agitation. Les groupes 2, 3, 5 et 8 ont les variances les plus faibles, avec des écarts-types trop faibles pour être représentés. Les groupes d'indices 1 et 6 présentent aussi une variance faible, mais avec un écart-type plus marqué que les groupes 2, 3, 5 et 8. Ces observations rappellent les interprétations des groupes de descripteurs d'accélération seuls faites en section 6.1.1 : le groupe *posé* pourrait être relié aux groupes 2, 3, 5 et 8 ; le groupe *calme* aux groupes 1 et 6 ; le groupe *agité* aux

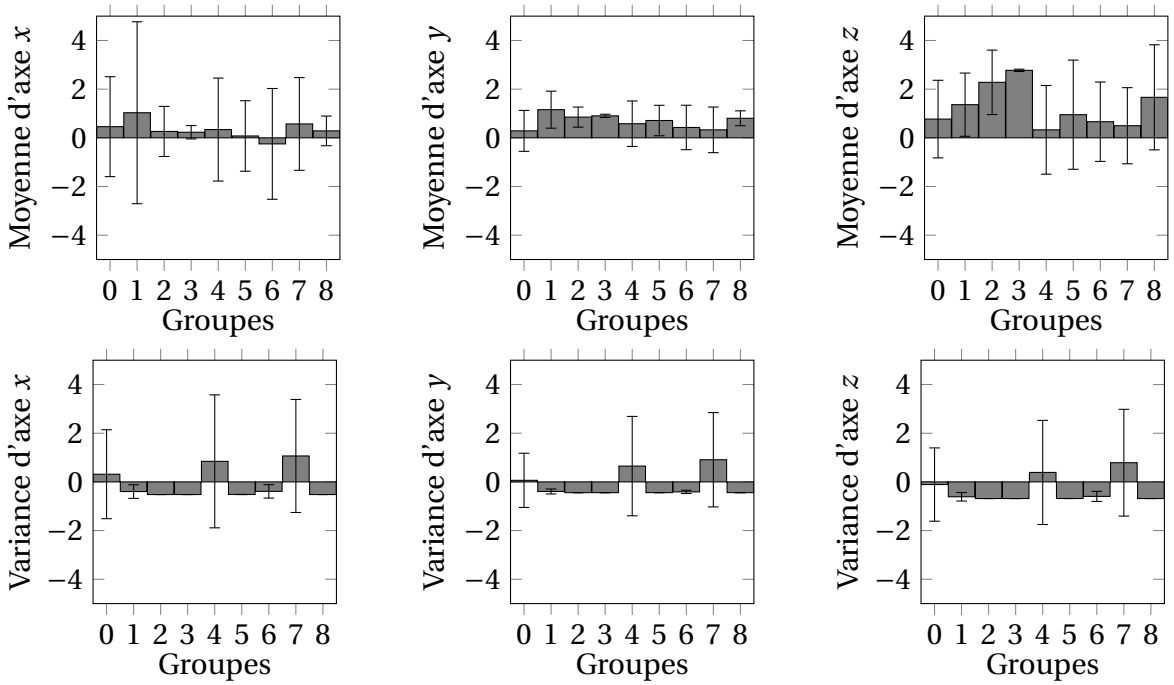


FIGURE 6.3: Descriptions des centroïdes des groupes suivant les moyennes et variances d'accélération sur les 3 axes

groupes 0, 4 et 7.

Concernant les moyennes d'accélération, on remarque la moyenne du groupe 3 sur l'axe z qui est la plus élevée, avec un écart-type faible. Par analogie avec les observations de la section 6.1.1, on peut supposer que le groupe représente l'immobilité du téléphone posé à plat sur une surface. La taille du groupe 3 est plus faible que celle du groupe *posé* de la section 6.1.1 (respectivement 46083 et 174782), ce qui laisse penser que beaucoup d'échantillons ont été affectés à d'autres groupes, probablement ceux d'indices 2, 5 et 8.

Ces observations permettent de compléter l'interprétation de la table 6.8. D'abord, l'association des groupes 2, 3, 5 et 8 à des faibles quantités de mouvement éclaire la forte présence de ces groupes dans la répartition des scènes *domicile*, *réunion*, *bureau* et *pause*. En particulier, si le groupe 3 représente le groupe *posé* de la section 6.1.1, alors il n'est pas surprenant qu'il se retrouve essentiellement dans les scènes du *domicile* et du *bureau*. La *rue* est principalement décrite par les groupes 0 et 7 représentatifs de l'agitation, ce qui est conforme à l'interprétation de l'activité principale de marche. Dans la section 6.1.1, nous avons suggéré que le groupe *calme* rassemble des situations de posture assise. Si l'on associe les groupes 1 et 6 au groupe *calme*, alors la forte présence de ces groupes dans les transports ainsi qu'au *magasin*, au *restaurant* et en *pause* est vraisemblable.

Les nuances qui apparaissent au sein des trois ensembles peuvent spécifier des activités ou des ambiances sonores particulières. C'est pourquoi, comme dans la section 6.1.2, nous nous intéressons aux descripteurs acoustiques dans les scènes de transport. Nous faisons l'hypothèse de la présence d'une forte énergie dans les coefficients acoustiques des basses fréquences qui serait associée aux bruits des véhicules motorisés. Dans la figure 6.4,

nous proposons d'observer l'histogramme des descripteurs acoustiques d'indices 1 et 2 pour les vecteurs des scènes de *bus* et de *voiture* du groupe 1. Pour rappel, les coefficients d'indices 1 et 2 représentent l'énergie respectivement dans les bandes de fréquences $[0;129Hz]$ et $[43;172Hz]$. L'énergie des coefficients de la scène de *voiture* atteint des valeurs très élevées. L'énergie des deux coefficients pour la scène du *bus* est plus faible, mais atteint quand même la valeur de 100 pour le coefficient d'indice 1. Ces observations sont en conformité avec l'hypothèse émise.

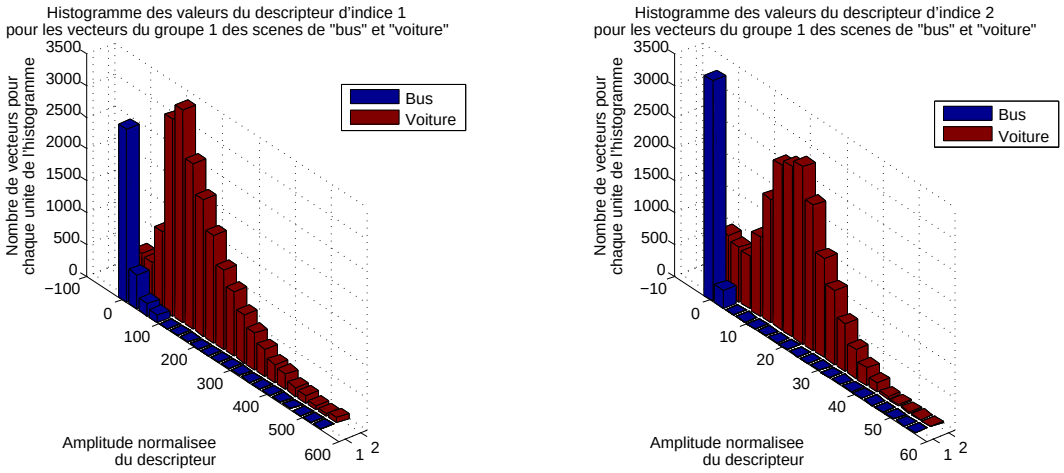


FIGURE 6.4: Histogrammes des vecteurs du groupe 1 pour les scènes du *bus* et de la *voiture* pour les descripteurs d'indices 1 (à gauche) et 2 (à droite)

Enfin, nous présentons la table 6.9 qui rapporte les résultats de classification opérée sur les étiquettes des groupes obtenus. On observe les fortes valeurs de la diagonale qui indiquent la bonne performance de classification des instances des groupes. La même conclusion qu'à l'issue des études précédentes des sections 6.1.1 et 6.1.2 s'impose : des groupes homogènes émergent mais leur interprétation nécessite des annotations que nous n'avons pas et la notion de scène ne peut être complètement retrouvée à partir des données.

g 0	g 4	g 7	g 1	g 6	g 2	g 3	g 5	g 8	← classé comme
99,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,9	g 0
0,0	**88,0**	9,8	0,2	1,9	0,0	0,0	0,1	0,0	g 4
0,8	2,6	**90,2**	0,5	4,3	0,0	0,0	1,6	0,0	g 7
4,7	0,1	0,8	**89,5**	1,5	1,6	0,3	1,4	0,1	g 1
0,0	0,2	3,7	1,3	**89,4**	0,0	0,0	5,5	0,0	g 6
0,0	0,0	0,0	1,1	0,0	**89,3**	5,9	3,6	0,1	g 2
0,0	0,0	0,0	0,1	0,0	1,9	**97,6**	0,0	0,4	g 3
0,0	0,0	1,5	1,4	3,8	3,0	0,0	**90,3**	0,0	g 5
0,4	0,0	0,0	0,0	0,0	0,0	0,0	0,0	**99,6**	g 8

TABLE 6.9: Matrice de confusion de la classification des groupes de vecteurs acoustiques et d'accélération

6.2 Approche par combinaison

Nous présentons dans cette section les résultats d'une dernière expérimentation de reconnaissance de scènes. L'approche consiste à reconnaître les éléments qui composent une scène puis à les combiner pour estimer la scène la plus vraisemblable. Cette approche repose sur la définition d'une scène par composition d'un lieu et d'une action, les deux éléments que nous avons identifiés dans la définition du chapitre 4. La stratégie de combinaison s'appuie sur la théorie de la fusion d'évidence de Dempster-Shafer. En effet, il n'y a pas de relation univoque entre les scènes et les éléments intermédiaires de lieu et d'action ; par conséquent, la présence d'incertitude dans les relations est propice à l'utilisation de cette théorie.

L'approche par composition représente une solution alternative au système de classification présenté dans le chapitre précédent. En outre, des informations intermédiaires sur la scène (telles que le lieu ou l'action) peuvent être estimées et compléter la description de la scène fournie. La solution que nous proposons tire aussi profit des résultats de la section 6.1 pour construire le module de reconnaissance d'action.

La section commence par présenter les expérimentations de reconnaissance d'action et de lieu dont les classifieurs sont employés dans le système. Ensuite, nous décrivons la solution de reconnaissance de scène, d'abord de manière théorique avec les notions de la fusion d'évidence de Dempster-Shafer ; puis de manière pratique en présentant l'application des concepts à notre réalisation.

6.2.1 Expérimentation de reconnaissance d'activité physique

Nous rapportons les travaux d'une étude que nous avons menée sur la reconnaissance d'activité physique d'une personne et du contexte du smartphone (Blachon et coll. (2014a)). Les objectifs étaient multiples :

— évaluer la pertinence du microphone en comparaison avec l'accéléromètre pour la reconnaissance de l'activité physique ;

— évaluer la capacité à reconnaître la position du smartphone sur la personne *via* une méthode d'apprentissage supervisé ;

— déterminer l'apport de la connaissance de la position du smartphone dans la reconnaissance de l'activité physique en intégrant cette connaissance dans le vecteur de descripteurs.

Concernant les concepts à reconnaître, nous avons considéré des activités physiques simples (suivant le sens donné dans la section 2.2.3) au cours desquelles les personnes peuvent porter le smartphone : la marche, la montée et descente d'escaliers, le saut et la course. Nous avons aussi considéré des attitudes immobiles en position debout, assise ou couchée. Ces attitudes sont observables dans de nombreuses situations (par exemple : les transports, le déjeuner, le travail sur bureau). Enfin, nous avons considéré le cas où le téléphone est posé sur une surface plane.

Le contexte du smartphone a déjà été évoqué dans la section 2.2.4 où l'on a notamment décrit son impact sur les mesures effectuées et potentiellement sur la tâche de reconnaissance. Dans les travaux de l'état de l'art, la prise en compte du contexte est limitée à la position et à l'orientation. Nous avons proposé une représentation plus vaste (Blachon et coll. (2014a)), incluant la quantité de mouvement de l'appareil, l'usage et la position. Chacun des trois éléments est décrit par des valeurs nominatives d'un ensemble fini. La quantité de mouvement peut être nulle, faible ou forte ; l'usage est décrit par deux valeurs (smartphone utilisé ou non) ; les positions considérées sont le sac, la poche du pantalon ou la main.

Le corpus de données qui a servi à l'expérimentation est décrit dans la section 3.4.2 du chapitre 3. Brièvement, 19 volontaires ont été équipés avec plusieurs smartphones situés dans les trois positions évoquées. Ils ont réalisé une séquence d'actions décrites dans des scénarios et supervisée par un expérimentateur. Le corpus exploitable est constitué de 408 minutes (un peu moins de 7 heures) et représentatif de 16 volontaires. Les données sont annotées avec les activités et attitudes mentionnées précédemment, ainsi qu'avec les positions du smartphone.

Le corpus de données a permis d'extraire un ensemble de descripteurs d'accélérations et d'ambiance sonore. Les descripteurs acoustiques sont calculés sur des fenêtres de 1024 échantillons puis moyennés sur une période équivalente de 2 secondes. Le calcul est similaire à ce qui a été décrit précédemment : ce sont des coefficients d'énergie de 40 filtres linéaires sur une échelle Mel, de bande équivalente à $[0;22050Hz]$. Les descripteurs d'accélération sont calculés sur des fenêtres de 2 secondes (synchronisées avec les fenêtres acoustiques moyennées). Les descripteurs incluent des mesures statistiques (moyenne, variance et énergie de la norme d'accélération, variance des accélération des 3 axes) et spectrales (coefficients d'énergie des bandes à 3 et 4 Hz). Un vecteur représente l'agrégation des descripteurs des deux sources, sur une fenêtre de 2 secondes.

Grâce au corpus de vecteurs annotés, nous avons mis en place plusieurs expérimentations de classification pour évaluer les trois hypothèses présentées précédemment. Le classifieur est entraîné et évalué suivant la méthode de validation croisée à 10 sous-ensembles avec répartition uniforme des activités annotées dans chaque sous-ensemble. Les classifieurs C4.5 et forêt d'arbres décisionnels sont employés dans les expérimentations. Le C4.5 a été entraîné avec élagage et un minimum de 100 vecteurs par feuille. La forêt d'arbres décisionnels est composée de 50 arbres de décision. Nous avons utilisé l'outil Weka pour l'expérimentation.

Nous présentons dans la figure 6.5 les résultats des expérimentations pour la forêt d'arbres décisionnels qui a obtenu les meilleures performances. Le diagramme de gauche illustre la f-mesure calculée pour la forêt d'arbres décisionnels dans la tâche de reconnaissance d'activité physique, suivant trois configurations. La configuration de *référence* représente le cas où la position du smartphone est inconnue. La configuration intitulée *vérité terrain* représente le cas où l'information de position est intégrée au vecteur de descripteurs ; dans chaque vecteur, on ajoute la vraie position de l'appareil. Dans la configuration appelée

inférence, un classifieur intermédiaire est entraîné pour reconnaître la position du smartphone. La prédiction est intégrée au vecteur des descripteurs pour la tâche de reconnaissance d'activité physique. Enfin, le graphique de droite illustre la f-mesure calculée pour la tâche de reconnaissance de la position du smartphone, évaluée en validation croisée à 10 sous-ensembles.

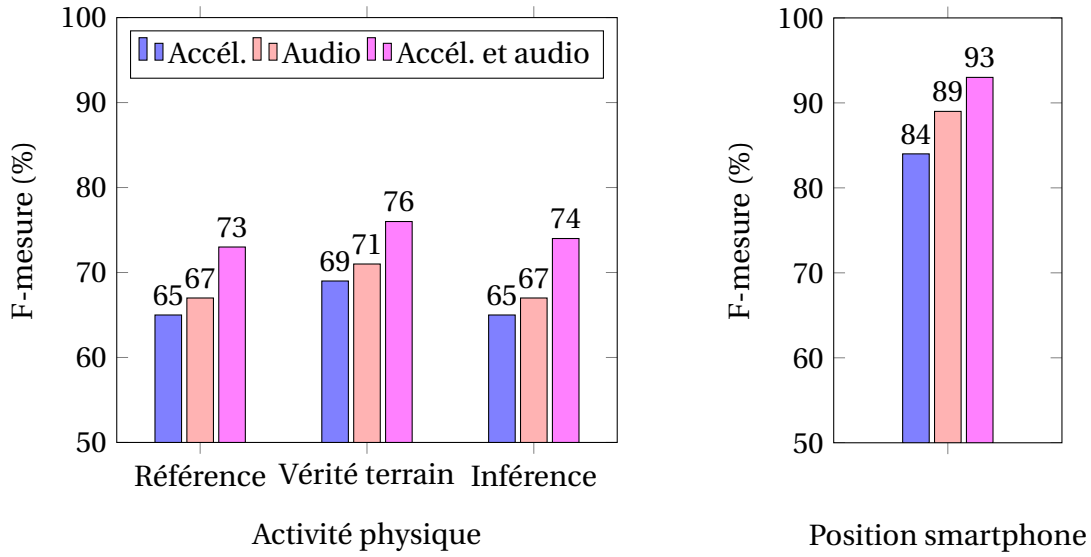


FIGURE 6.5: F-mesures calculées pour la reconnaissance d'activités physiques et de position du smartphone, en validation croisée à 10 sous-ensembles avec le classifieur de forêt d'arbres décisionnels (RF)

Concernant l'objectif d'évaluation de la pertinence des données acoustiques dans la tâche de reconnaissance d'activité physique, on constate que pour les trois configurations, la f-mesure est plus élevée avec les descripteurs acoustiques qu'avec les descripteurs d'accélération. En outre, la combinaison des deux sources permet d'obtenir des résultats encore plus élevés que lorsqu'une seule des deux sources est employée.

La comparaison des résultats des configurations *vérité terrain* et *référence* indique que l'intégration de la position du smartphone dans le vecteur de descripteurs à évaluer permet d'augmenter la performance de 3 à 4 points suivant les ensembles de descripteurs considérés. La configuration *vérité terrain* représente une estimation théorique de la performance car l'information de la position du smartphone est connue avec certitude. En comparaison, l'expérimentation où la position est inférée ne permet pas d'atteindre des performances similaires. Pour expliquer ce résultat, nous décrivons d'abord le graphique de droite de la figure 6.5 qui indique les valeurs de f-mesure pour la classification de position du smartphone. Les valeurs sont très élevées, quelle que soit la configuration de descripteurs choisis. Ces valeurs sont obtenues suivant une validation croisée à 10 sous-ensembles. Dans le cadre de l'expérimentation d'*inférence*, le classifieur de position du smartphone a été entraîné sur une sous-partie du corpus, ce qui a réduit le nombre d'exemples considéré pendant l'entraînement et probablement affecté la reconnaissance de la position. Par suite, l'attribut de position a pu être considéré comme peu pertinent dans l'apprentissage, ce qui pourrait expliquer les résultats très similaires à la *référence*.

Les résultats sont encourageants pour la réalisation de modules dédiés à la reconnaissance de l'activité physique et de la position du smartphone. Cependant, afin d'intégrer ces modules dans un système composite de reconnaissance de scènes, les vecteurs de scènes doivent être annotés avec les informations d'activité physique et de position du smartphone. Nous n'avons pas ces informations, c'est pourquoi nous proposons une solution alternative dans la section suivante.

6.2.2 Expérimentation de reconnaissance d'agitation

Dans la section 6.1.1, nous avons présenté trois groupes homogènes obtenus après regroupement de vecteurs composés de descripteurs d'accélération. L'interprétation des centroïdes des groupes a permis d'émettre des hypothèses sur l'agitation, représentées par les trois étiquettes *posé*, *calme* et *agité*. Nous avons aussi effectué une expérimentation de classification des vecteurs annotés avec les étiquettes issues du regroupement. La matrice de confusion a été présentée (voir table 6.5), affichant des taux très élevés de reconnaissance, ce qui a permis de conclure sur l'homogénéité et la différentiation possible des groupes.

Puisque les données de scènes ne sont pas annotées avec les activités physiques, nous proposons d'utiliser les trois étiquettes précédentes pour entraîner un module de reconnaissance d'agitation de la personne. Nous rapportons dans la table 6.10 les résultats de reconnaissance de l'expérimentation décrite dans la section 6.1.1, exprimés par les mesures de rappel et précision.

TABLE 6.10: Rappel et précision moyens calculés pour la tâche de reconnaissance des groupes d'agitation

Rappel moy. (%)	**Précision moy. (%)**
99,4	99,7

Pour rappel, les trois groupes sont obtenus par regroupement non-supervisé. Nous avons remarqué dans la section 6.1.1 que ces groupes sont homogènes et identifiables par une méthode de classification. Le rappel moyen correspond à la moyenne du rappel pour chacun des trois groupes. Il en va de même pour la précision moyenne. Des valeurs aussi élevées indiquent que le classifieur identifie très bien les étiquettes des vecteurs de chaque groupe et commet peu d'erreurs dans la prédiction.

6.2.3 Expérimentation de reconnaissance du lieu

Concernant la reconnaissance du lieu, nous utilisons les étiquettes proposées dans le chapitre 4 et associées aux scènes de la manière suivante :

— *intérieur privé* : *domicile* ;

— *intérieur public* : *restaurant, magasin* ;

— *intérieur professionnel* : *bureau, réunion, pause* ;

— *extérieur* : *rue* ;

— *transports* : *voiture, bus, tramway, train, bateau.*

L'expérimentation nécessite l'annotation des vecteurs de scènes avec les étiquettes de lieux correspondants. La classification est effectuée avec une forêt d'arbres décisionnels, suivant une validation croisée à 10 sous-ensembles. Nous choisissons ce classifieur car il a présenté les meilleures performances dans les expérimentations du chapitre 5. De plus, lors de l'évaluation d'un vecteur, le vote effectué parmi les arbres qui le composent peut être assimilé à une probabilité. L'usage de probabilités est pertinent pour la combinaison des prédictions des modules, comme nous le verrons dans le section 6.2.4.

Nous présentons dans la table 6.11 les mesures de rappel et précision de la forêt d'arbres décisionnels pour la tâche. Les valeurs sont élevées, ce qui n'est pas surprenant pour la tâche (nous renvoyons le lecteur aux matrices de confusion de la figure 5.6 du chapitre 5 pour la tâche de classification de macro-environnements).

TABLE 6.11: Performances de reconnaissance du lieu en validation croisée à 10 sous-ensembles

Rappel (%)	**Précision (%)**
94,4	94,4

6.2.4 Combinaison des éléments par fusion

Pour la combinaison des prédictions des classifieurs dits intermédiaires, notre solution repose sur la théorie de fusion de preuves (*evidential fusion theory* en anglais) de Dempster-Shafer (Dempster (1968), Shafer et coll. (1976)) et, notamment, les réseaux de fusion de preuves pour l'inférence de concepts. Nous nous inspirons de l'application de cette théorie à la reconnaissance d'activités physiques (Hong et coll. (2009)) et à la reconnaissance de chutes (Aguilar et coll. 2014). Dans ces articles, un réseau de fusion de preuves est mis en place pour évaluer une hypothèse (une activité ou une chute par exemple). L'hypothèse est décrite par la composition de sous-activités, elles-mêmes associées aux observations effectuées. L'ensemble forme un réseau représenté par un graphe dont les nœuds sont des variables et les arcs, les relations permises entre les nœuds. En outre, une fonction de masse (au sens de la théorie de Dempster-Shafer) est associée à chaque nœud du graphe. La fonction de masse quantifie la croyance des relations entre les valeurs possibles du nœud et celles de ses nœuds parents.

Dans notre solution, nous proposons l'usage d'un tel réseau. Nous décrivons d'abord les concepts théoriques de la fusion d'évidence de Dempster-Shafer. Puis nous abordons notre réalisation de fusion pour la reconnaissance de scène.

6.2.4.1 Éléments théoriques sur la fusion d'évidence de Dempster-Shafer

Nous définissons quelques concepts théoriques nécessaires à la compréhension et à l'application de la théorie de fusion d'évidence de Dempster-Shafer. Cette théorie complète l'approche bayésienne par la considération de l'incertitude dans le calcul de probabilités. Pour

illustrer ce concept, Dempster utilise l'exemple suivant dans un article de 1968 (Dempster (1968)). Considérons une carte géographique recouverte d'eau et de terre. La carte est connue à 80 %. La terre est présente sur 70 % de la zone connue et l'eau sur les 30 % restants. La partie inconnue peut être remplie d'eau ou de terre, dans des proportions inconnues. La question est de savoir quelle est la probabilité qu'un point sélectionné aléatoirement sur la carte entière soit recouvert d'eau. Les différentes valeurs permettent de dire avec certitude que la probabilité est supérieure à 24 % (30 % multipliés par 80 %) et inférieure à 44 % (le complément des 56 % de la partie terrestre certaine). La valeur 0,24 peut être associée à l'étiquette *eau*, la valeur 0,56 à l'étiquette *terre* et la valeur 0,20 (le complément des deux probabilités précédentes) aux deux étiquettes *eau* et *terre*.

Définitions

On appelle **ensemble de discernement** Θ (*frame of discernment* en anglais) l'ensemble exhaustif des valeurs possibles mutuellement exclusives d'une variable. Dans l'exemple, il s'agit des étiquettes *eau* et *terre*.

À partir de cet ensemble, on définit l'**ensemble des parties d'un ensemble** 2^{Θ} (*power set*) qui décrit la totalité des sous-ensembles qu'il est possible de construire à partir des éléments de l'ensemble de discernement. Trois parties sont considérées dans l'exemple : les singletons *eau* et *terre* ainsi que la pair (*eau,terre*). L'ensemble vide $\emptyset$ est aussi considéré comme une partie. La théorie de Dempster-Shafer se distingue de l'approche bayésienne en considérant la réalisation possible de plusieurs valeurs simultanément.

Dans la théorie de Dempster-Shafer (Dempster (1968), Shafer et coll. (1976)), on considère la croyance de réalisation d'une partie de Θ. La croyance est exprimée par une **fonction de masse** (*mass function* en anglais), définie sur l'ensemble des parties de Θ et associant une valeur comprise entre 0 et 1 à chacun de ses éléments. En outre, la fonction vérifie les conditions suivantes (Hong et coll. 2009) :

$$\begin{cases} m(\emptyset) = 0 & \emptyset : \text{l'ensemble vide} \\ \sum\limits_{A \subseteq \Theta} m(A) = 1 & \text{A : une partie de } \Theta \end{cases} \tag{6.3}$$

Les deux propriétés sont vérifiées dans l'exemple précédent. L'application de la fonction de masse à un sous-ensemble de plus d'un élément symbolise l'incertitude associée aux éléments du sous-ensemble telle que l'on ne peut préciser la croyance sur les éléments qui le composent.

Enfin, une hypothèse est évaluée par deux quantités qui expriment différentes fonctions de masse. La **croyance** représente la limite inférieure de vraisemblance associée à une hypothèse par la somme des fonctions de masse des concepts inclus dans l'hypothèse. La **plausibilité** (*plausibility*) représente la borne supérieure par la somme des fonctions de masse des concepts qui contiennent l'hypothèse. Les deux équations suivantes illustrent les définitions

pour une hypothèse A :

$$\text{bel}(A) = \sum_{B \subseteq A} m(B) \tag{6.4}$$

$$\text{pls}(A) = \sum_{B \supseteq A} m(B) \tag{6.5}$$

L'exemple précédent considère trois hypothèses : la présence d'eau ; la présence de terre ; ou l'incertitude, représentée par la présence possible d'eau et de terre. La croyance associée à la présence d'eau dans un point sélectionné aléatoirement sur la carte est égale à 0,24. La plausibilité considère en plus l'hypothèse de la présence possible d'eau et de terre, associée à 0,20. Donc la plausibilité associée à la présence d'eau est de 0,44.

Propriétés des relations

Association de parties (*multivalued mapping*) : lorsque plusieurs sources d'information proposent des preuves exprimées dans différents ensembles de discernement, il est nécessaire d'établir des relations pour les associer. La relation par association de parties indique les paires d'éléments de Θ_E et Θ_H qui peuvent être simultanément vraies (Strat 1987) :

$$\Theta_{E,H} \subseteq \Theta_E \times \Theta_H \tag{6.6}$$

Association de parties par évidence (*evidential mapping* (Hong et coll. 2009) : lorsque l'on dispose d'informations supplémentaires sur la relation entre les parties de deux ensembles, on peut compléter l'association en faisant correspondre à un élément e_i de l'ensemble d'origine une partie $H_{i,j}$ de l'ensemble de destination et la fonction de masse associée f :

$$\begin{aligned} \Gamma^* \quad : \quad e_i \quad \rightarrow \quad & \Big\{\big(H_{i,j}, f(e_i \rightarrow H_{i,j})\big), \ldots, \big(H_{i,n}, f(e_i \rightarrow H_{i,n})\big)\Big\} \\ & \text{avec } e_i \in \Theta_E,\ H_{i,j} \in 2^{\Theta_H},\ i = 1 \ldots \text{card}(\Theta_E) \text{ et } j = 1 \ldots n \end{aligned} \tag{6.7}$$

La fonction f représente la croyance qu'un élément e_i de Θ_E puisse être vrai en même temps qu'une partie $H_{i,j}$ de 2^{Θ_H}. Un élément e_i peut être associé à plusieurs parties. La croyance, exprimée entre 0 et 1, se répartit parmi les parties de l'ensemble de destination associées. Ces propriétés s'expriment par les équations suivantes (Hong et coll. 2009) :

$$\begin{aligned} H_{i,j} &\neq \emptyset \quad j = 1, \ldots, n \\ f(e_i \rightarrow H_{i,j}) &> 0 \quad j = 1, \ldots, n \\ \sum_{j=1}^{n} f(e_i \rightarrow H_{i,j}) &= 1 \end{aligned} \tag{6.8}$$

Propriétés de combinaison des inférences

Combinaison des fonctions de masse : Lorsque plusieurs sources indépendantes fournissent des preuves associées à des croyances, il est possible de les combiner suivant la règle

de combinaison de fonctions de masses proposée par Dempster :

$$m(C) = \frac{\sum_{A\cap B=C} m_1(A)m_2(B)}{1-\sum_{A\cap B=\emptyset} m_1(A)m_2(B)} = \frac{\sum_{A\cap B=C} m_1(A)m_2(B)}{\sum_{A\cap B\neq\emptyset} m_1(A)m_2(B)} \tag{6.9}$$

Combinaison par somme uniformément pondérée : Lorsque les sources ne sont pas indépendantes, la règle précédente ne peut être appliquée. Hong et coll. (2009) proposent un opérateur de somme uniformément pondérée (adapté de l'opérateur de somme proposé par McClean et Scotney (1997)) :

$$m(A) = m_1 \oplus \cdots \oplus m_N(A) = \frac{1}{N}\sum_{i=1}^{N} m_i(A) \tag{6.10}$$

6.2.4.2 Description de l'expérimentation

Notre approche considère l'estimation de la scène la plus vraisemblable par la fusion des preuves fournies par deux classifieurs de lieu et d'action. Nous faisons l'hypothèse que l'utilisateur se trouve dans l'une des scènes considérées. Nous évaluons la vraisemblance que l'utilisateur se trouve dans chacune des scènes. Notre système est représenté par un graphe dans la figure 6.6. Les nœuds de *lieu* et d'*action* représentent les classifieurs ; le nœud de la *pause* représente l'une des scènes à évaluer. Les deux arcs indiquent une relation entre chacun des classifieurs et l'hypothèse de scène, que nous détaillons plus loin.

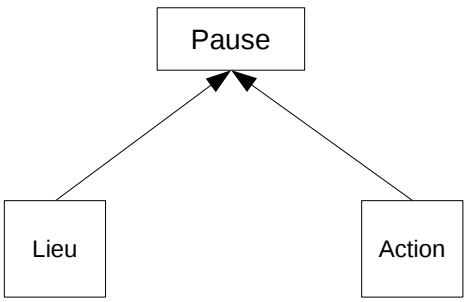


FIGURE 6.6: Réseau de preuves pour la reconnaissance de scène par fusion intermédiaire

Nous définissons d'abord les *ensembles de discernement* des trois nœuds dans la table 6.12. Les scènes considérées proviennent des annotations obtenues dans le corpus. Les éléments de lieux sont repris de la section 6.2.3 et les éléments d'action de la section 6.2.2. Les sous-ensembles que peuvent considérer les deux classifieurs sont limités aux singletons des éléments car les classifieurs sont probabilistes et considèrent les éléments indépendamment. Les sous-ensembles de scènes à considérer sont également les singletons des éléments car nous considérons les scènes individuellement.

Les relations entre les deux classifieurs et le nœud de la scène sont décrites par des *associations de parties par évidence* (définies dans la section 6.2.4.1). Pour un élément de l'ensemble de discernement des classifeurs de lieu et d'action, les sous-ensembles de scènes qui peuvent être vraies sont associées avec leur fonction de masse dans les tables 6.13 et 6.14 Les

TABLE 6.12: Illustration des ensembles de valeurs mutuellement exclusives pour quelques noeuds du réseau

Noeud	Type	Ensemble des valeurs mutuellement exclusives
Scène	Scène	{Bateau, Bus, Voiture, Domicile, Réunion, Pause, Bureau, Restaurant, Magasin, Rue, Train, Tramway}
Action	Classifieur	{Posé, Calme, Agité}
Lieu	Classifieur	{Int. priv., Int. pro., Int. pub., Extérieur, Transports}

croyances indiquées dans les tables sont calculées de manière empirique à partir de statistiques du corpus de données de l'expérimentation. L'association des lieux considère d'abord les scènes possibles pour chaque lieu. Par exemple, seules les scènes du *tramway*, du *bus*, du *train*, du *bateau* et de la *voiture* peuvent être associées aux *transports*. De plus, par l'hypothèse qu'une seule scène peut se produire à la fois, nous considérons les sous-ensembles des scènes composées d'une scène individuelle. Le calcul de la croyance est effectué par le rapport du nombre de vecteurs d'observation de la scène sur le nombre de vecteurs d'observations du lieu. Par exemple, l'association des *transports* au *tramway* a une croyance de $14914/62895 = 0,237$. Le calcul des croyances des actions vers les scènes est similaire. Dans ce cas, le dénominateur représente le nombre de vecteurs du groupe associé, déterminé dans la table 6.3 de la section 6.1.1

Les classifieurs réalisent des prédictions sous forme de probabilités associées à chacun des éléments qu'ils considèrent (lieu ou action). Ces probabilités sont converties en la croyance de réalisation de l'élément associé. Comme précédemment, les seuls sous-ensembles considérés sont les singletons des éléments.

TABLE 6.13: Association des lieux aux scènes*

Lieu	Scène	Instances	Association
Transports	Tramway	14914	{transports} → {({tramway}, 0,237)}
Transports	Bus	13703	{transports} → {({bus}, 0,218)}
Transports	Train	3780	{transports} → {({train}, 0,060)}
Transports	Bateau	1393	{transports} → {({bateau}, 0,022)}
Transports	Voiture	29015	{transports} → {({voiture}, 0,462)}
Intérieur pro.	Réunion	47660	{int. pro.} → {({réunion}, 0,294)}
Intérieur pro.	Pause	9068	{int. pro.} → {({pause}, 0,056)}
Intérieur pro.	Bureau	105378	{int. pro.} → {({bureau}, 0,650)}
Intérieur pub.	Magasin	10876	{int. pub.} → {({magasin}, 0,207)}
Intérieur pub.	Restaurant	41760	{int. pub.} → {({restaurant}, 0,793)}
Extérieur	Rue	30146	{extérieur} → {({rue}, 1,0)}
Intérieur privé	Domicile	83619	{int. priv.} → {({domicile}, 1,0)}

* Les associations impossibles, telles que la scène du domicile dans les transports, ont une fonction de masse nulle et ne sont pas écrites.

La prédiction des classifieurs est pondérée par une mesure de confiance qui résulte du calcul de la précision des classifieurs. Nous reprenons les mesures de précision des tables 6.10 et 6.11 respectivement de 99,7 % pour l'action et de 94,5 % pour le lieu. Ces valeurs sont obtenues sur le même corpus que celui qui sert à l'évaluation du système de reconnaissance par fusion. Il y a donc un biais expérimental, dont nous sommes conscients. Cependant, nous ne disposons pas d'un corpus suffisamment important pour évaluer la précision des classifieurs sur une autre partie.

TABLE 6.14: Association des actions aux scènes (les statistiques nulles ne sont pas représentées)

Action	Scène	Instances	Association
agité	magasin	5424	{agité} → {({magasin},0.046)}
agité	bureau	8239	{agité} → {({bureau},0.070)}
agité	tramway	5027	{agité} → {({tramway},0.043)}
agité	bus	8305	{agité} → {({bus},0.071)}
agité	train	1640	{agité} → {({train},0.014)}
agité	réunion	6549	{agité} → {({réunion},0.056)}
agité	rue	25438	{agité} → {({rue},0.217)}
agité	restaurant	14268	{agité} → {({restaurant},0.121)}
agité	voiture	27251	{agité} → {({voiture},0.232)}
agité	domicile	11745	{agité} → {({domicile},0.100)}
agité	pause	3275	{agité} → {({pause},0.028)}
agité	bateau	287	{agité} → {({bateau},0.002)}
posé	magasin	316	{posé} → {({magasin},0.002)}
posé	bureau	77318	{posé} → {({bureau},0.493)}
posé	tramway	108	{posé} → {({tramway},0.001)}
posé	réunion	24484	{posé} → {({réunion},0.156)}
posé	train	485	{posé} → {({train},0.003)}
posé	restaurant	7039	{posé} → {({restaurant},0.045)}
posé	domicile	45532	{posé} → {({domicile},0.290)}
posé	pause	1507	{posé} → {({pause},0.010)}
calme	magasin	5136	{calme} → {({magasin},0.044)}
calme	bureau	19821	{calme} → {({bureau},0.169)}
calme	tramway	9779	{calme} → {({tramway},0.084)}
calme	bus	5398	{calme} → {({bus},0.046)}
calme	train	1655	{calme} → {({train},0.014)}
calme	réunion	16627	{calme} → {({réunion},0.142)}
calme	rue	4708	{calme} → {({rue},0.040)}
calme	restaurant	20453	{calme} → {({restaurant},0.175)}
calme	voiture	1764	{calme} → {({voiture},0.015)}
calme	domicile	26342	{calme} → {({domicile},0.225)}
calme	pause	4286	{calme} → {({pause},0.037)}
calme	bateau	1106	{calme} → {({bateau},0.009)}

Pour terminer la description du réseau, la combinaison des croyances d'une scène suivant les différents "chemins" de lieux et d'actions est calculée en appliquant l'*opérateur de somme uniformément pondérée* (défini dans la section 6.2.4.1) sur les différentes croyances issues de chaque configuration de lieux et actions. Cet opérateur est préféré à la règle de combinaison de Dempster car les sources ne sont pas indépendantes (en effet, les classifieurs fonctionnent sur des descripteurs communs). Dans la suite, nous illustrons le processus de calcul global par un exemple de reconnaissance de la scène du *domicile*.

Le calcul est composé de trois étapes :

1. calcul de la masse associée à chaque lieu et action, par combinaison de la prédiction des classifieurs avec la confiance ;
2. calcul de la masse associée à une scène sachant un lieu ou une action par application de la relation d'association de parties par évidence ;
3. combinaison des masses d'une même scène pour les différents lieux et actions évalués.

Nous considérons les prédictions fournies par les deux classifieurs à partir d'un vecteur de descripteur. Les équations 6.11 et 6.12 présentent le calcul de la première étape. Les coefficients c_{lieu} et c_{action} représentent la mesure de confiance des classifieurs de lieu et d'action.

$$\begin{cases} m(\{\text{int. priv.}\}) = p(\text{lieu} = \text{int.priv.}) \times c_{\text{lieu}} = 0,77 \times 0,945 = 0,728 \\ m(\{\text{int. pro.}\}) = p(\text{lieu} = \text{int. pro.}) \times c_{\text{lieu}} = 0,05 \times 0,945 = 0,047 \\ m(\{\text{int. pub.}\}) = p(\text{lieu} = \text{int.pub.}) \times c_{\text{lieu}} = 0,02 \times 0,945 = 0,019 \\ m(\{\text{extérieur}\}) = p(\text{lieu} = \text{extérieur}) \times c_{\text{lieu}} = 0,16 \times 0,945 = 0,151 \\ m(\{\text{transports}\}) = p(\text{lieu} = \text{transports}) \times c_{\text{lieu}} = 0,0 \times 0,945 = 0,0 \end{cases} \tag{6.11}$$

$$\begin{cases} m(\{\text{posé}\}) = p(\text{action} = \text{posé}) \times c_{\text{action}} = 0,0 \times 0,997 = 0,0 \\ m(\{\text{calme}\}) = p(\text{action} = \text{calme}) \times c_{\text{action}} = 0,0 \times 0,997 = 0,0 \\ m(\{\text{agité}\}) = p(\text{action} = \text{agité}) \times c_{\text{action}} = 1,0 \times 0,997 = 0,997 \end{cases} \tag{6.12}$$

Dans un second temps, la croyance de la scène du *domicile* (considérée pour l'exemple) est calculée conditionnellement aux différents lieux et actions. Les relations et valeurs sont indiquées dans les tables 6.13 et 6.14. L'équation 6.13 présente l'ensemble des calculs. Par les associations décrites dans la table 6.13, le seul lieu qui puisse être associé au *domicile* est l'*intérieur privé*. Cela justifie les croyances nulles résultantes pour les autres lieux. De plus, nous avons vu dans l'équation 6.12 que la prédiction du classifieur est certaine pour l'action *agité*, laissant une croyance nulle pour les deux autres actions.

La troisième étape consiste à calculer la masse de la scène du *domicile* par application de l'*opérateur de somme uniformément pondérée* de l'équation 6.10. L'équation 6.14 présente le calcul.

En effectuant un calcul identique pour les autres scènes, on obtient les croyances résumées dans la table 6.15. La scène la plus vraisemblable est celle qui affiche la croyance la plus élevée, il s'agit du domicile.

$$\begin{cases} m_{\text{int. priv.}}(\{\text{domicile}\}) = m(\{\text{int. priv.}\}) \times m(\{\text{int. priv.}\} \rightarrow \{\text{domicile}\}) \\ \qquad = 0,728 \times 1,0 = 0,728 \\ m_{\text{int. pro.}}(\{\text{domicile}\}) = m(\{\text{int. pro.}\}) \times m(\{\text{int. pro.}\} \rightarrow \{\text{domicile}\}) \\ \qquad = 0,047 \times 0,0 = 0,0 \\ m_{\text{int. pub.}}(\{\text{domicile}\}) = m(\{\text{int. pub.}\}) \times m(\{\text{int. pub.}\} \rightarrow \{\text{domicile}\}) \\ \qquad = 0,019 \times 0,0 = 0,0 \\ m_{\text{extérieur}}(\{\text{domicile}\}) = m(\{\text{extérieur}\}) \times m(\{\text{extérieur}\} \rightarrow \{\text{domicile}\}) \\ \qquad = 0,151 \times 0,0 = 0,0 \\ m_{\text{transports}}(\{\text{domicile}\}) = m(\{\text{transports}\}) \times m(\{\text{transports}\} \rightarrow \{\text{domicile}\}) \\ \qquad = 0,0 \times 0,0 = 0,0 \\ m_{\text{posé}}(\{\text{domicile}\}) = m(\{\text{posé}\}) \times m(\{\text{posé}\} \rightarrow \{\text{domicile}\}) \\ \qquad = 0,0 \times 0,290 = 0,0 \\ m_{\text{agité}}(\{\text{domicile}\}) = m(\{\text{agité}\}) \times m(\{\text{agité}\} \rightarrow \{\text{domicile}\}) \\ \qquad = 0,997 \times 0,100 = 0,100 \\ m_{\text{calme}}(\{\text{domicile}\}) = m(\{\text{calme}\}) \times m(\{\text{calme}\} \rightarrow \{\text{domicile}\}) \\ \qquad = 0,0 \times 0,225 = 0,0 \end{cases} \tag{6.13}$$

$$\begin{aligned} m(\{\text{domicile}\}) &= \frac{1}{N}\sum_{e\in\Theta} m_e(\{\text{domicile}\}) \quad \text{avec } \Theta = \Theta_{\text{lieu}} \cup \Theta_{\text{action}} \text{ et } N = \text{card}(\Theta) \\ &= \frac{1}{5+3}\big(0,728+0,100\big) \\ &= 0,104 \end{aligned} \tag{6.14}$$

TABLE 6.15: Croyances des différentes scènes après intégration des croyances intermédiaires des lieux et actions

	Bateau	Bus	Voiture	Domicile	Réunion	Pause	Bureau	Restau.	Magasin	Rue	Train	Tramway
Croyance	0,0003	0,009	0,029	**0,104**	0,009	0,004	0,013	0,017	0,006	0,046	0,002	0,005

6.2.4.3 Résultats de la fusion et commentaires

Nous avons effectué l'expérimentation en validation croisée sur 10 sous-ensembles. Le taux de classification est de 71,5 %. Pour rappel, le taux de classification de la forêt d'arbres décisionnels dans la même configuration de sources et en validation croisée stratifiée à 10 sous-ensembles est de 90,3 % (voir section 5.3.1). La première observation consiste à dire que le système de reconnaissance par fusion n'est pas aussi bon que le système de classification du chapitre 5.

Nous présentons les mesures de rappel et de précision pour chacune des scènes dans la table 6.16. On remarque une grande disparité dans la reconnaissance des scènes avec le *bateau*, le *bus*, la *réunion*, la *pause*, le *magasin* et le *train* qui ne sont jamais identifiés. À l'inverse, les autres scènes sont souvent reconnues et prédites par le classifieur. Nous expliquons cette observation par les associations de lieux et de scènes, qui sont très restrictives, avec des croyances très marquées. Par exemple, l'*intérieur professionnel* est associé au *bureau* avec une croyance de 0,650, très forte relativement aux associations avec la *pause* et de la *réunion*, respectivement de 0,056 et 0,294. Ainsi, la combinaison de la prédiction du lieu avec les associations de lieux et de scènes favorise la scène la plus vraisemblable pour le lieu à la probabilité la plus élevée.

Les associations d'actions aux scènes ne mettent pas plus en valeur les scènes "faibles". L'action dite *agité* est associée au *bureau* suivant une croyance de 0,070, à la *réunion* avec 0,056 et à la *pause* avec 0,028. L'action dite *posé* est associée au *bureau* avec une croyance de 0,493. Enfin, l'action dite *calme* est associée au *bureau* avec 0,169, à la *réunion* avec 0,142 et à la *pause* avec 0,037.

La même observation est possible pour les scènes du *restaurant* et du *magasin*. Concernant les *transports*, l'action dite *calme* permet la distinction entre le *tramway* et la *voiture* (qui est le transport le plus représenté) avec des croyances respectives de 0,084 et 0,015. La croyance dans le *bus* est aussi plus élevée que dans la *voiture* (0,046). Cela justifie pourquoi deux scènes des transports sont reconnues pendant la fusion (contrairement aux autre environnements où la scène avec la plus grande croyance est la seule proposée).

Malgré le biais mis en évidence vers les scènes les plus fréquentes, le résultat de cette

TABLE 6.16: Mesures de rappel et précision des scènes

	Bateau	Bus	Voiture	Domicile	Réunion	Pause	Bureau	Restau.	Magasin	Rue	Train	Tramway
Rappel	0,0	0,0	94,1	93,6	0,0	0,0	96,0	85,9	0,0	90,0	0,0	44,1
Précision	0,0	0,0	60,5	94,0	0,0	0,0	63,9	63,9	0,0	77,2	0,0	49,6

expérimentation est encourageant. Le système pourrait être enrichi de nouveaux modules et la stratégie de Dempster-Shafer permettrait d'intégrer des sources aux formats différents. Ainsi, une évolution possible serait de prendre en compte des sources événementielles du fonctionnement de l'appareil telles que les informations d'applications utilisées ou l'état de l'écran (allumé ou éteint) et de les intégrer sous forme de module. La modélisation de ces sources peut être heuristique (par exemple, si l'identifiant de la borne Wi-Fi de mon domicile est reconnue, alors il y a 95 % de chance que je sois chez moi, à l'intérieur).

Le fonctionnement par modules permet de fournir des informations différentes sur la scène. Certes, l'information d'environnement peut être directement déduite de la scène suivant notre modélisation. Mais l'information d'action est complémentaire au lieu pour estimer l'action de la personne. Avec l'ajout d'une source supplémentaire comme l'interaction de l'utilisateur avec l'appareil, on peut imaginer différencier les actions dues à l'interaction de celles simplement dues au mouvement ou au déplacement de la personne. Cette approche hiérarchique permet de réduire la dépendance au matériel grâce à la multiplication et la diversité des sources d'information. Par exemple, si une source est indisponible temporairement, l'estimation des hypothèses de scènes est possible grâce aux autres sources disponibles. Un autre avantage de cette solution est l'adaptation des poids à la personne (par exemple, par un apprentissage spécifique).

6.3 Bilan du chapitre

La première conclusion du chapitre porte sur la complexité des scènes considérées et perçues par les humains relativement aux observations issues de mesures physiques. L'étude menée dans la section 6.1 n'a pas permis d'aboutir à une interprétation très poussée de la scène. Cependant, nous avons montré la cohérence des groupes issus du regroupement non-supervisé, ce qui justifie de vouloir interpréter les groupes. L'étude des ambiances sonores présentée est aussi très limitée par le manque d'annotations sur les ambiances considérées.

La seconde conclusion porte sur la représentation de la scène qui est très simple pour le moment. Les éléments de lieu et d'action sont nécessaires mais insuffisants pour la réalisation d'un système de reconnaissance composite. La mise en évidence d'éléments supplémentaires pourrait permettre d'améliorer le système de reconnaissance par fusion d'évidences. Celui-ci présente des avantages relativement aux objectifs industriels, mais reste limité par le nombre de sources et de modules de représentation employés. La théorie de fusion ne peut être pleinement appliquée car la considération de classifieurs empêche la représentation de l'incertitude. Ainsi, l'étude d'autres sources de données pour découvrir d'autres motifs de la composition d'une scène devrait permettre une meilleure compréhen-

sion et une meilleure représentation du système.

CHAPITRE 7

Conclusion

7.1 Bilan

Le manuscrit rapporte le travail de thèse réalisé dans l'objectif de construire un système embarqué sur un smartphone capable de reconnaître la scène de l'utilisateur à partir des sources de données disponibles. Très tôt au cours du doctorat, nous avons considéré que, pour atteindre cet objectif, un ensemble d'objectifs intermédiaires devait être atteint. Ceux-ci sont la conséquence de contraintes scientifiques telles que le manque de connaissance sur le concept de scène ; ou de contraintes pratiques comme l'absence d'un corpus de données correspondant aux critères fixés ; d'autres sont des objectifs industriels tels que le souhait de pouvoir décrire une scène suivant plusieurs niveaux d'abstraction. Le travail décrit s'est efforcé de remplir ces objectifs ; les contributions sont les suivantes.

La première contribution de la thèse est notre système de classification, présenté au chapitre 5, qui répond au problème principal de l'identification des scènes. Le système a été évalué dans des conditions réalistes (les données proviennent de situations réelles) suivant une méthodologie qui simule aussi un cas réaliste d'apprentissage centré sur un utilisateur et tenant compte de vecteurs de données acquis dans un passé proche. En outre, l'évaluation considère plusieurs configurations de capteurs et de descripteurs de données ainsi que le déséquilibre de la représentation des scènes dans l'apprentissage. Dans le cas d'une validation croisée stratifiée à 10 sous-ensembles, sur un corpus composé de données d'un seul volontaire, la forêt d'arbres décisionnels (RF) a obtenu le meilleur rappel de classification avec la valeur de 90,3 %. L'arbre de décision C4.5 présente aussi un résultat très proche. Les autres classifieurs sont moins satisfaisants. Ces résultats constituent une référence de classification et confirment la possibilité de reconnaître directement une scène, dans des conditions d'expérimentation et, notamment, si toutes les sources de données évaluées sont présentes.

Pour parvenir à ce résultat de classification, nous avons fait le choix d'une approche par apprentissage automatique supervisé qui nécessite des données annotées. L'appareil visé par le sujet (le smartphone) nous a incité à faire l'acquisition d'un corpus de données collecté sur un appareil du genre. En outre, le concept souhaité (la scène de la vie quotidienne) a orienté le choix vers des données réelles, collectées *in vivo*. D'autres contraintes d'annotations et de sources de données ont mené à une recherche infructueuse de corpus existants. Suite à cela, nous avons fait le choix d'effectuer notre propre collecte de données. Plusieurs contraintes ont été rencontrées lors de l'établissement du protocole de collecte de données.

Le procédé d'annotation a dû être renforcé pour vérifier les annotations renseignées à la volée par les volontaires. La sécurité des données est un autre problème, qu'il a fallu traiter lors des différentes étapes d'acquisition, de transfert et de stockage. La gestion de la vie privée et l'anonymat des données ont aussi représenté une contrainte, gérée par des mesures de protection sur toute la chaîne de traitement, à commencer par l'enregistrement sur l'appareil. Finalement, deux collectes ont été effectuées, annotées et exploitées, dont la principale est celle portant sur les scènes. Elle totalise plus de 500 heures de données réparties dans 80 enregistrements uniques ; plus d'une vingtaine de volontaires ont participé, dont 6 qui ont collecté des scènes dans leur vie quotidienne ; plus de dix smartphones de différentes marques et gammes ont servi aux collectes ; et des lieux essentiellement locaux (la région de Grenoble) mais aussi d'autres régions de France, d'Irlande, de Hongrie et de Singapour. L'ensemble du corpus collecté et annoté ainsi que l'application de collecte RECORDME sont deux contributions de la thèse et répondent à l'objectif d'acquisition d'un corpus.

Le travail de thèse décrit dans le manuscrit s'inscrit dans un contexte où le concept de scène et les situations visées par l'application industrielle sont mal connues. En effet, d'une part, les situations visées dans l'application finale par le partenaire industriel sont peu décrites ; d'autre part, la notion de scène est floue dans l'état de l'art et les travaux effectués l'abordent en ne considérant que certains des éléments qui la composent (par exemple : reconnaissance du lieu, de l'activité, de l'ambiance sonore). Pour parvenir à l'objectif principal de reconnaissance de scène, nous avons considéré nécessaire d'approfondir la compréhension du concept de scène. Nous proposons une définition dans le chapitre 4 qui résulte de l'étude de travaux existants et des annotations du corpus. L'étude combinée de la notion de contexte, proche de la scène, et des travaux de reconnaissance effectués a permis de mettre en évidence la notion de composition. Les éléments de lieu et d'action sont apparus comme nécessaires. L'étude des annotations a confirmé leur importance mais a aussi montré les limites de la description. La définition que nous proposons est générale, ce qui permet de considérer de nombreuses situations, mais elle manque encore de précision. Elle représente une première approximation qui peut être exploitée pour la modélisation d'une scène, mais son imprécision requiert encore du travail pour améliorer la compréhension.

Une autre contribution de la thèse est la confirmation de la pertinence de l'usage du microphone, en complément de l'accéléromètre. L'usage de ce dernier a été montré dans les tâches de reconnaissance d'activité physique, de reconnaissance de la position du smartphone ainsi que dans des activités humaines plus complexes, qui peuvent être associées à des lieux ou des scènes (par exemple, la préparation du repas se réalise la plupart du temps dans une cuisine). Nous avons vu dans l'état de l'art que le microphone est pertinent pour la reconnaissance d'ambiance sonore telle que celle de lieux particuliers. Plusieurs résultats décrits dans le manuscrit confirment la pertinence de l'usage du microphone. La sélection d'attributs effectuée dans le chapitre 5 a retenu 10 coefficients d'énergie acoustique, localisés dans une bande de fréquence de 0 à 1077 Hz. L'expérimentation de reconnaissance d'activité physique et de position du smartphone décrite dans la section 6.2.1 du chapitre 6 a montré que l'usage du microphone peut amener à des résultats équivalents à celui de l'ac-

céléromètre. Les résultats indiquent aussi que l'usage commun des deux sources dépasse les résultats obtenus pour les sources individuelles.

Nous avons exploité le modèle de scène en présentant une solution de reconnaissance alternative dans le chapitre 6. Il s'agit d'un système composé de modules (des classifieurs) dédiés à la reconnaissance des éléments de lieu et d'action qui caractérisent les scènes considérées. Le module procède à une combinaison des prédictions des deux classifieurs suivant la théorie de fusion de Dempster-Shafer. Les prédictions sont transformées en fonctions de masse associées aux différentes valeurs de variables considérées. Les essais effectués avec le système ne sont pas concluants pour le moment mais nous proposons plusieurs explications à cela. D'abord, comme cela a déjà été dit, le modèle de scène est encore très général et imprécis, ce qui peut avoir une influence sur la description des scènes. En outre, l'usage de classifieurs impose de considérer des probabilités, ce qui ne permet pas de tirer profit de la représentation d'incertitude et de la considération de réalisations multiples simultanées proposées par la théorie de Dempster-Shafer. En outre, cette théorie permet d'unifier des représentation continues et symboliques, ce qui permettrait de prendre en compte les sources de données du fonctionnement du téléphone pour compléter les modules du système. Dans l'état actuel, ce système ne représente pas une solution aboutie, mais il ouvre des pistes de recherche au-delà de la thèse.

7.2 Perspectives

Pour aller plus loin, nous suggérons d'approfondir la compréhension de la notion de scène. D'un point de vue théorique, nous avons distingué la scène de l'utilisateur du contexte du smartphone. En outre, la scène est composée d'un lieu et d'une action. Ces deux éléments sont identifiés par classification et ne sont pas décrits précisément. Les résultats de classification sur ces éléments sont bons, ce qui laisse imaginer que l'association directe de lieu ou de l'action aux données est suffisante. Mais il est impossible de généraliser, pour le moment, à un plus grand nombre d'utilisateurs. Le travail doit donc porter sur une meilleure caractérisation de la scène et du contexte du smartphone. Pour cela, nous disposons d'un ensemble de données (annotées ou non) qui peuvent être exploitées. L'analyse non-supervisée des données d'accélération et d'ambiances sonores a mis en avant des groupes homogènes et identifiables, dont l'interprétation est limitée par manque d'annotations. De telles études peuvent être reproduites à partir des données collectées par les autres capteurs. De plus, nous suggérons d'étudier les données issues du fonctionnement de l'appareil. En effet, des informations telles que l'état d'allumage de l'écran, les applications utilisées ou les appels passés peuvent donner lieu à des hypothèses simples sur l'usage du smartphone et ainsi servir d'annotations pour l'interprétation des groupes non-supervisés.

Nous n'avons pas représenté l'évolution du temps dans les solutions proposées. Seule l'expérimentation de détection de transitions considère une séquence de vecteurs, ordonnée de manière chronologique. Les systèmes proposés jusqu'ici ne permettent pas d'intégrer des connaissances sur l'évolution dynamique des scènes. Or, l'analyse du corpus de don-

nées collecté a montré que les scènes peuvent être très longues avec une durée minimum de plusieurs minutes. La représentation de cette dynamique et l'intégration au système de reconnaissance pourraient être étudiées. En outre, si l'on considère l'usage de l'application industrielle dans un contexte d'activités routinières, des hypothèses plus précises de durée peuvent être formulées.

Dans l'évaluation des solutions proposées dans le chapitre 5, nous avons implicitement émis l'hypothèse que toutes les scènes rencontrées sont connues. Cette hypothèse est irréaliste ; une solution serait de rendre le système capable de considérer l'incertitude ou le doute comme un état. Comme nous l'avons déjà écrit, la théorie de Dempster-Shafer considère l'incertitude (qui s'exprime par la réalisation simultanée de plusieurs valeurs d'une variable). Ainsi, le système de combinaison basé sur cette théorie pourrait permettre la considération de l'incertitude et donc doter le système de la capacité d'exprimer le doute dans la prédiction. En combinant la prise en compte de la séquence avec celle de l'incertitude, il est possible d'envisager l'apprentissage d'une nouvelle scène. Le doute prolongé pendant une période pourrait servir à sauvegarder un sous-ensemble des vecteurs de la période. Par la suite, on peut imaginer que le système suggère à l'utilisateur la détection d'une nouvelle scène et demande une étiquette, ce qui permettrait d'initier la modélisation d'une nouvelle classe.

Pour conclure, les contributions de la thèse ne se limitent pas aux outils, connaissance, corpus et résultats décrits précédemment. Ensemble, elles s'organisent au sein d'une démarche suivie pour répondre au problème initial de reconnaissance de scène embarquée sur smartphone. Cette démarche est, elle-même, le résultat de l'expression de plusieurs contraintes : l'application finale peu détaillée, le manque de connaissance sur la notion de scène, le choix d'une approche supervisée et le besoin d'un corpus annoté, l'usage du smartphone et ses contraintes matérielles. Finalement, la méthodologie suivie apparaît aussi comme une contribution du travail car elle propose des solutions et ouvre des perspectives de recherche dans d'autres domaines.

Bibliographie

AHARONY, N., PAN, W., IP, C., KHAYAL, I. et PENTLAND, A. (2011). Social fmri : Investigating and shaping social mechanisms in the real world. *Pervasive and Mobile Computing*, 7(6): 643–659.

AINSWORTH, B. E., HASKELL, W. L., HERRMANN, S. D., MECKES, N., BASSETT JR, D. R., TUDOR-LOCKE, C., GREER, J. L., VEZINA, J., WHITT-GLOVER, M. C. et LEON, A. S. (2011). 2011 compendium of physical activities : a second update of codes and met values. *Medicine and science in sports and exercise*, 43(8):1575–1581.

AVCI, A., BOSCH, S., MARIN-PERIANU, M., MARIN-PERIANU, R. et HAVINGA, P. (2010). Activity recognition using inertial sensing for healthcare, wellbeing and sports applications : A survey. *In Architecture of computing systems (ARCS), 2010 23rd international conference on*, pages 1–10. VDE.

AZIZYAN, M., CONSTANDACHE, I. et ROY CHOUDHURY, R. (2009). Surroundsense : mobile phone localization via ambience fingerprinting. *In Proceedings of the 15th annual international conference on Mobile computing and networking*, pages 261–272. ACM.

BAO, L. et INTILLE, S. S. (2004). Activity recognition from user-annotated acceleration data. *In Pervasive computing*, pages 1–17. Springer.

BENGIO, Y. (2009). Learning deep architectures for ai. *Foundations and trends® in Machine Learning*, 2(1):1–127.

BILMES, J. A. *et al.* (1998). A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. *International Computer Science Institute*, 4(510):126.

BLACHON, D., COŞKUN, D. et PORTET, F. (2014a). On-line context aware physical activity recognition from the accelerometer and audio sensors of smartphones. *In Ambient Intelligence*, pages 205–220. Springer.

BLACHON, D., PORTET, F., BESACIER, L. et TASSART, S. (2014b). Recordme : A smartphone application for experimental collections of large amount of data respecting volunteer's privacy. *In Ubiquitous Computing and Ambient Intelligence. Personalisation and User Adapted Services*, pages 345–348. Springer.

BOUTEN, C. V., KOEKKOEK, K. T., VERDUIN, M., KODDE, R. et JANSSEN, J. D. (1997). A triaxial accelerometer and portable data processing unit for the assessment of daily physical activity. *Biomedical Engineering, IEEE Transactions on*, 44(3):136–147.

BREIMAN, L. (1996). Bagging predictors. *Machine learning*, 24(2):123–140.

BREIMAN, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

BROWN, P. J., BOVEY, J. D. et CHEN, X. (1997). Context-aware applications : from the laboratory to the marketplace. *Personal Communications, IEEE*, 4(5):58–64.

BRUNETTE, W., SUNDT, M., DELL, N., CHAUDHRI, R., BREIT, N. et BORRIELLO, G. (2013). Open data kit 2.0 : expanding and refining information services for developing regions. *In Proceedings of the 14th Workshop on Mobile Computing Systems and Applications*, page 10. ACM.

CARROLL, A. et HEISER, G. (2013). The systems hacker's guide to the galaxy energy usage in a modern smartphone. *In Proceedings of the 4th Asia-Pacific Workshop on Systems*, page 5. ACM.

CAVALCANTE AGUILAR, P. A., BOUDY, J., ISTRATE, D., DORIZZI, B. et MOURA MOTA, J. C. (2014). A dynamic evidential network for fall detection. *Biomedical and Health Informatics, IEEE Journal of*, 18(4):1103–1113.

CHEN, G., KOTZ, D. *et al.* (2000). A survey of context-aware mobile computing research. Rapport technique, Technical Report TR2000-381, Dept. of Computer Science, Dartmouth College.

CLARKSON, B., SAWHNEY, N. et PENTLAND, A. (1998). Auditory context awareness via wearable computing. *Energy*, 400(600):20.

COPPOLA, P., DELLA MEA, V., DI GASPERO, L., MENEGON, D., MISCHIS, D., MIZZARO, S., SCAGNETTO, I. et VASSENA, L. (2010). The context-aware browser. *Intelligent Systems, IEEE*, 25(1):38–47.

COUTAZ, J., CROWLEY, J. L., DOBSON, S. et GARLAN, D. (2005). Context is key. *Communications of the ACM*, 48(3):49–53.

CROWLEY, J. L., COUTAZ, J., REY, G. et REIGNIER, P. (2002). Perceptual components for context aware computing. *In UbiComp 2002 : Ubiquitous Computing*, pages 117–134. Springer.

CVETKOVIĆ, B., KALUŽA, B., MILIĆ, R. et LUŠTREK, M. (2013). Towards human energy expenditure estimation using smart phone inertial sensors. *In Ambient Intelligence*, pages 94–108. Springer.

DASH, M. et LIU, H. (1997). Feature selection for classification. *Intelligent Data Analysis*, 1:131–156.

DEMPSTER, A. P. (1968). A generalization of bayesian inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 205–247.

DERNBACH, S., DAS, B., KRISHNAN, N. C., THOMAS, B. L. et COOK, D. J. (2012). Simple and complex activity recognition through smart phones. *In Intelligent Environments (IE), 2012 8th International Conference on*, pages 214–221. IEEE.

DEY, A. K. (1998). Context-aware computing : The cyberdesk project. *In Proceedings of the AAAI 1998 Spring Symposium on Intelligent Environments*, pages 51–54.

DEY, A. K., ABOWD, G. D. et SALBER, D. (2001). A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications. *Human-computer interaction*, 16(2):97–166.

DIACONITA, I., REINHARDT, A., ENGLERT, F., CHRISTIN, D. et STEINMETZ, R. (2014). Do you hear what i hear? using acoustic probing to detect smartphone locations. *In Pervasive Computing and Communications Workshops (PERCOM Workshops), 2014 IEEE International Conference on*, pages 1–9. IEEE.

EAGLE, N. et PENTLAND, A. (2006). Reality mining : sensing complex social systems. *Personal and ubiquitous computing*, 10(4):255–268.

FRANKLIN, D. et FLASCHBART, J. (1998). All gadget and no representation makes jack a dull environment. *In Proceedings of the AAAI 1998 Spring Symposium on Intelligent Environments*, pages 155–160.

HALL, M., FRANK, E., HOLMES, G., PFAHRINGER, B., REUTEMANN, P. et WITTEN, I. H. (2009). The weka data mining software : an update. *ACM SIGKDD explorations newsletter*, 11(1): 10–18.

HALL, M. A. (1999). *Correlation-based feature selection for machine learning*. Thèse de doctorat, The University of Waikato.

HALL, M. A. et HOLMES, G. (2003). Benchmarking attribute selection techniques for discrete class data mining. *Knowledge and Data Engineering, IEEE Transactions on*, 15(6):1437–1447.

HANSON, A. R. et RISEMAN, E. M. (1978). Visions : A computer system for interpreting scenes. *Computer vision systems*, 78:303–334.

HARTUNG, C., LERER, A., ANOKWA, Y., TSENG, C., BRUNETTE, W. et BORRIELLO, G. (2010). Open data kit : tools to build information services for developing regions. *In Proceedings of the 4th ACM/IEEE International Conference on Information and Communication Technologies and Development*, page 18. ACM.

HINTON, G., DENG, L., YU, D., DAHL, G. E., MOHAMED, A.-r., JAITLY, N., SENIOR, A., VANHOUCKE, V., NGUYEN, P., SAINATH, T. N. *et al.* (2012). Deep neural networks for acoustic modeling in speech recognition : The shared views of four research groups. *Signal Processing Magazine, IEEE*, 29(6):82–97.

HONG, X., NUGENT, C., MULVENNA, M., MCCLEAN, S., SCOTNEY, B. et DEVLIN, S. (2009). Evidential fusion of sensor data for activity recognition in smart homes. *Pervasive and Mobile Computing*, 5(3):236–252.

HOSEINI-TABATABAEI, S. A., GLUHAK, A. et TAFAZOLLI, R. (2013). A survey on smartphone-based systems for opportunistic user context recognition. *ACM Computing Surveys (CSUR)*, 45(3):27.

INCEL, O. D., KOSE, M. et ERSOY, C. (2013). A review and taxonomy of activity recognition on mobile phones. *BioNanoScience*, 3(2):145–171.

JACQUET, C., BELLIK, Y. et BOURDA, Y. (2004). *A context-aware locomotion assistance device for the blind*. Springer.

KERN, N., SCHIELE, B. et SCHMIDT, A. (2007). Recognizing context for annotating a live life recording. *Personal and Ubiquitous Computing*, 11(4):251–263.

KIUKKONEN, N., BLOM, J., DOUSSE, O., GATICA-PEREZ, D. et LAURILA, J. (2010). Towards rich mobile phone datasets : Lausanne data collection campaign. *Proc. ICPS, Berlin*.

KWAPISZ, J. R., WEISS, G. M. et MOORE, S. A. (2011). Activity recognition using cell phone accelerometers. *ACM SigKDD Explorations Newsletter*, 12(2):74–82.

LANE, N. D., MILUZZO, E., LU, H., PEEBLES, D., CHOUDHURY, T. et CAMPBELL, A. T. (2010). A survey of mobile phone sensing. *Communications Magazine, IEEE*, 48(9):140–150.

LARCHER, A., BONASTRE, J.-F., FAUVE, B. G., LEE, K.-A., LÉVY, C., LI, H., MASON, J. S. et PARFAIT, J.-Y. (2013). Alize 3.0-open source toolkit for state-of-the-art speaker recognition. *In INTERSPEECH*, pages 2768–2772.

LE, V. B., MELLA, O., FOHR, D. *et al.* (2007). Speaker diarization using normalized cross likelihood ratio. *In INTERSPEECH*, volume 7, pages 1869–1872.

LESTER, J., CHOUDHURY, T. et BORRIELLO, G. (2006). A practical approach to recognizing physical activities. *In Pervasive Computing*, pages 1–16. Springer.

LIU, H. et YU, L. (2005). Toward integrating feature selection algorithms for classification and clustering. *Knowledge and Data Engineering, IEEE Transactions on*, 17(4):491–502.

LU, H., YANG, J., LIU, Z., LANE, N. D., CHOUDHURY, T. et CAMPBELL, A. T. (2010). The jigsaw continuous sensing engine for mobile phone applications. *In Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems*, pages 71–84. ACM.

MA, L., SMITH, D. et MILNER, B. P. (2003). Context awareness using environmental noise classification. *In INTERSPEECH*.

MARMASSE, N. et SCHMANDT, C. (2000). Location-aware information delivery withcommotion. *In Handheld and Ubiquitous Computing*, pages 157–171. Springer.

MCCLEAN, S. et SCOTNEY, B. (1997). Using evidence theory for the integration of distributed databases. *International Journal of Intelligent Systems*, 12(10):763–776.

MIAO, Y. (2014). Kaldi+ pdnn : building dnn-based asr systems with kaldi and pdnn. *arXiv preprint arXiv :1401.6984*.

MILUZZO, E., LANE, N. D., FODOR, K., PETERSON, R., LU, H., MUSOLESI, M., EISENMAN, S. B., ZHENG, X. et CAMPBELL, A. T. (2008). Sensing meets mobile social networks : the design, implementation and evaluation of the cenceme application. *In Proceedings of the 6th ACM conference on Embedded network sensor systems*, pages 337–350. ACM.

MILUZZO, E., PAPANDREA, M., LANE, N. D., LU, H. et CAMPBELL, A. T. (2010). Pocket, bag, hand, etc.-automatically detecting phone context through discovery. *Proc. PhoneSense 2010*, pages 21–25.

MINSKY, M. (1975). A framework for representing knowledge. *The Psychology of Computer Vision*.

PARK, J.-g., PATEL, A., CURTIS, D., TELLER, S. et LEDLIE, J. (2012). Online pose classification and walking speed estimation using handheld devices. *In Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 113–122. ACM.

PEARL, J. (2011). Bayesian networks. https://escholarship.org/uc/item/53n4f34m. [Online ; accessed 4-mai-2015].

PELTONEN, V., TUOMI, J., KLAPURI, A., HUOPANIEMI, J. et SORSA, T. (2002). Computational auditory scene recognition. *In Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, volume 2, pages II–1941. IEEE.

QUINLAN, J. R. (1993). C4. 5 : Programs for machine learning.

RAVI, N., SCOTT, J., HAN, L. et IFTODE, L. (2008). Context-aware battery management for mobile phones. *In Pervasive Computing and Communications, 2008. PerCom 2008. Sixth Annual IEEE International Conference on*, pages 224–233. IEEE.

REDDY, S., MUN, M., BURKE, J., ESTRIN, D., HANSEN, M. et SRIVASTAVA, M. (2010). Using mobile phones to determine transportation modes. *ACM Transactions on Sensor Networks (TOSN)*, 6(2):13.

SCHANK, R. et ABELSON, R. (1977). Scripts, plans, goals and understanding.

SCHILIT, B., ADAMS, N. et WANT, R. (1994). Context-aware computing applications. *In Mobile Computing Systems and Applications, 1994. WMCSA 1994. First Workshop on*, pages 85–90. IEEE.

SHAFER, G. *et al.* (1976). *A mathematical theory of evidence*, volume 1. Princeton university press Princeton.

SHOAIB, M., BOSCH, S., INCEL, O. D., SCHOLTEN, H. et HAVINGA, P. J. (2014). Fusion of smartphone motion sensors for physical activity recognition. *Sensors*, 14(6):10146–10176.

SIIRTOLA, P. et RONING, J. (2013). Ready-to-use activity recognition for smartphones. *In Computational Intelligence and Data Mining (CIDM), 2013 IEEE Symposium on*, pages 59–64. IEEE.

SOHN, T., VARSHAVSKY, A., LAMARCA, A., CHEN, M. Y., CHOUDHURY, T., SMITH, I., CONSOLVO, S., HIGHTOWER, J., GRISWOLD, W. G. et DE LARA, E. (2006). Mobility detection using everyday gsm traces. *In UbiComp 2006 : Ubiquitous Computing*, pages 212–224. Springer.

STRAT, T. M. (1987). The generation of explanations within evidential reasoning systems. *In IJCAI*, volume 87, pages 1097–1104.

VINCIARELLI, A., PANTIC, M. et BOURLARD, H. (2009). Social signal processing : Survey of an emerging domain. *Image and Vision Computing*, 27(12):1743–1759.

WAGNER, D. T., RICE, A. et BERESFORD, A. R. (2014). Device analyzer : Understanding smartphone usage. *In Mobile and Ubiquitous Systems : Computing, Networking, and Services*, pages 195–208. Springer.

WARD, A., JONES, A. et HOPPER, A. (1997). A new location technique for the active office. *Personal Communications, IEEE*, 4(5):42–47.

WIKI, O. (2014). Main page — openstreetmap wiki,. http://wiki.openstreetmap.org/w/index.php?title=Main_Page&oldid=1060762. [Online ; accessed 4-février-2015].

WINOGRAD, T. (2001). Architectures for context. *Human-Computer Interaction*, 16(2):401–419.

WITTEN, I. H. et FRANK, E. (2005). *Data Mining : Practical machine learning tools and techniques*. Morgan Kaufmann.

YAN, Z., CHAKRABORTY, D., MISRA, A., JEUNG, H. et ABERER, K. (2012). Semantic activity classification using locomotive signatures from mobile phones. Rapport technique.

Bibliographie personnelle

[1] BLACHON, D., COŞKUN, D. ET PORTET, F. (2014). On-line Context Aware Physical Activity Recognition from the Accelerometer and Audio Sensors of Smartphones. Dans *Ambient Intelligence*, pages 205–220, Springer International Publishing

[2] BLACHON, D., PORTET, F., BESACIER, L., ET TASSART, S. (2014). RecordMe : A Smartphone Application for Experimental Collections of Large Amount of Data Respecting Volunteer's Privacy. Dans *Ubiquitous Computing and Ambient Intelligence. Personalisation and User Adapted Services* (pp. 345-348). Springer International Publishing.

Annexes

TABLE 7.1: Table de description des filtres acoustiques

	Indice	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
(Hz)	Début	0	43	129	172	258	345	431	560	646	775	904	1077	1206	1378	1593	1766	1981	2239	2498	2799
	Centre	43	129	172	258	345	431	560	646	775	904	1077	1206	1378	1593	1766	1981	2239	2498	2799	3101
	Fin	129	172	258	345	431	560	646	775	904	1077	1206	1378	1593	1766	1981	2239	2498	2799	3101	3445
	Indice	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
(Hz)	Début	3101	3445	3790	4221	4651	5125	5642	6202	6804	7494	8226	9001	9862	10810	11843	12920	14126	15461	16882	18475
	Centre	3445	3790	4221	4651	5125	5642	6202	6804	7494	8226	9001	9862	10810	11843	12920	14126	15461	16882	18475	20155
	Fin	3790	4221	4651	5125	5642	6202	6804	7494	8226	9001	9862	10810	11843	12920	14126	15461	16882	18475	20155	22050

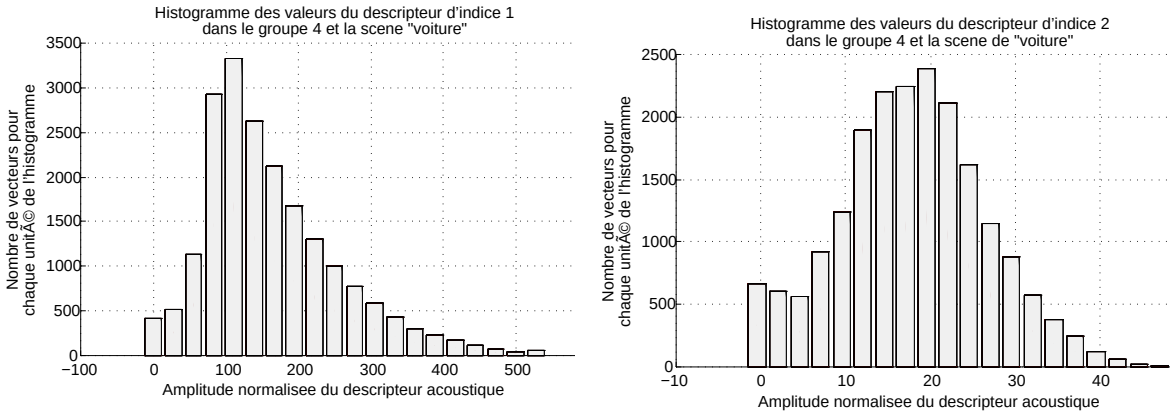


FIGURE 7.1: Histogramme des coefficients d'énergie d'indices 1 et 2 des vecteurs du groupe 4 et de la scène de *voiture*

TABLE 7.2: Résultat de la sélection des descripteurs par la méthode de ratio de gain d'information

Ratio	Descripteur	Ratio	Descripteur	Ratio	Descripteur
0.17076	Pressure_mean	0.10928	Gyro_energy_band_19_Hz	0.0768	MelFilterBank#38
0.15552	MelFilterBank#1	0.10872	Accel_energy_band_9_Hz	0.07174	Magnetic_energy_band_1_Hz
0.15541	Gyro_variance	0.10782	Accel_energy_band_8_Hz	0.05653	Magnetic_energy_band_24Hz
0.15409	MelFilterBank#0	0.10773	Gyro_mean	0.05628	Magnetic_energy_band_23_Hz
0.14785	MelFilterBank#2	0.10757	Gyro_energy_band_20_Hz	0.05587	Magnetic_energy_band_25Hz
0.14456	MelFilterBank#3	0.1074	Gyro_energy_band_16_Hz	0.05479	Magnetic_energy_band_4_Hz
0.14269	Gyro_energy_band_2_Hz	0.10666	Gyro_energy_band_25Hz	0.05463	Magnetic_energy_band_22_Hz
0.14245	MelFilterBank#6	0.10658	MelFilterBank#22	0.05363	Magnetic_energy_band_20_Hz
0.14166	MelFilterBank#7	0.10619	Accel_energy_band_13_Hz	0.05314	Magnetic_energy_band_3_Hz
0.14164	MelFilterBank#4	0.10605	Gyro_energy_band_23_Hz	0.05288	Magnetic_energy_band_19_Hz
0.14103	MelFilterBank#5	0.10568	Gyro_energy_band_17_Hz	0.05251	Magnetic_energy_band_16_Hz
0.14067	MelFilterBank#9	0.10567	Accel_energy_band_1_Hz	0.05203	Magnetic_energy_band_17_Hz
0.13931	MelFilterBank#8	0.10523	Gyro_energy_band_24Hz	0.05024	Magnetic_energy_band_21_Hz
0.13845	Gyro_energy	0.10512	Gyro_energy_band_22_Hz	0.04889	Magnetic_energy_band_13_Hz
0.1372	MelFilterBank#10	0.1048	Accel_energy_band_6_Hz	0.04824	Magnetic_energy_band_10_Hz
0.13625	MelFilterBank#11	0.10427	MelFilterBank#23	0.0478	Magnetic_energy_band_14_Hz
0.13525	MelFilterBank#12	0.10261	MelFilterBank#24	0.04671	Magnetic_energy_band_11_Hz
0.13284	Gyro_energy_band_3_Hz	0.10234	MelFilterBank#25	0.04636	Magnetic_energy_band_18_Hz
0.13198	MelFilterBank#13	0.10093	Accel_energy_band_18_Hz	0.04487	Magnetic_energy_band_7_Hz
0.12983	Gyro_energy_band_1_Hz	0.10056	Accel_energy_band_15_Hz	0.0427	Magnetic_energy_band_15_Hz
0.12863	Accel_energy_band_3_Hz	0.09999	Accel_energy_band_23_Hz	0.04188	Magnetic_energy_band_12_Hz
0.12827	MelFilterBank#14	0.0999	Accel_energy_band_21_Hz	0.04173	Magnetic_energy_band_8_Hz
0.12399	MelFilterBank#15	0.09958	Magnetic_mean	0.04123	Magnetic_energy_band_5_Hz
0.1237	Gyro_energy_band_6_Hz	0.09926	MelFilterBank#30	0.03594	Magnetic_energy_band_9_Hz
0.12356	Gyro_energy_band_5_Hz	0.09924	MelFilterBank#26	0.03477	Pressure_energy_band_24Hz
0.12315	Gyro_energy_band_7_Hz	0.09924	Accel_energy_band_22_Hz	0.03317	Pressure_variance
0.12258	Gyro_energy_band_4_Hz	0.09873	Accel_energy_band_16_Hz	0.03276	Pressure_energy_band_3_Hz
0.1199	Gyro_energy_band_8_Hz	0.09845	Accel_energy_band_24Hz	0.02925	Pressure_energy
0.11969	Accel_variance	0.09818	Accel_energy_band_20_Hz	0.02883	Magnetic_energy_band_6_Hz
0.11924	Gyro_energy_band_9_Hz	0.09788	Accel_energy_band_2_Hz	0.01869	Pressure_energy_band_23_Hz
0.11923	Gyro_energy_band_10_Hz	0.0972	MelFilterBank#27	0.01796	Pressure_energy_band_4_Hz
0.1183	MelFilterBank#17	0.09698	Accel_energy_band_14_Hz	0.01742	Pressure_energy_band_2_Hz
0.11805	MelFilterBank#16	0.0964	Accel_energy_band_25Hz	0.01738	Pressure_energy_band_25Hz
0.11707	MelFilterBank#19	0.09577	Accel_energy_band_12_Hz	0.01679	Pressure_energy_band_22_Hz
0.11578	Accel_energy	0.09448	Accel_energy_band_17_Hz	0.01586	Pressure_energy_band_5_Hz
0.11564	Gyro_energy_band_21_Hz	0.09396	Accel_energy_band_19_Hz	0.01114	Pressure_energy_band_7_Hz
0.11557	MelFilterBank#18	0.09374	MelFilterBank#35	0.01084	Pressure_energy_band_20_Hz
0.11543	MelFilterBank#20	0.09226	MelFilterBank#33	0.00913	Pressure_energy_band_19_Hz
0.11457	Accel_energy_band_4_Hz	0.09206	MelFilterBank#36	0.0089	Pressure_energy_band_8_Hz
0.1139	Gyro_energy_band_12_Hz	0.09193	MelFilterBank#32	0.00771	Pressure_energy_band_21_Hz
0.11373	MelFilterBank#21	0.09191	Magnetic_variance	0.00758	Pressure_energy_band_6_Hz
0.11262	Gyro_energy_band_11_Hz	0.09021	MelFilterBank#34	0.00651	Pressure_energy_band_9_Hz
0.11252	Gyro_energy_band_18_Hz	0.09007	MelFilterBank#28	0.00625	Pressure_energy_band_1_Hz
0.11237	Gyro_energy_band_13_Hz	0.08953	MelFilterBank#31	0.00599	Pressure_energy_band_11_Hz
0.11158	Gyro_energy_band_15_Hz	0.08823	MelFilterBank#29	0.00595	Pressure_energy_band_18_Hz
0.11062	Accel_energy_band_10_Hz	0.08745	MelFilterBank#37	0.00483	Pressure_energy_band_15_Hz
0.11015	Accel_energy_band_7_Hz	0.08535	Accel_mean	0.00461	Pressure_energy_band_12_Hz
0.11012	Accel_energy_band_5_Hz	0.08506	Magnetic_energy_band_2_Hz	0.00436	Pressure_energy_band_14_Hz
0.10985	Accel_energy_band_11_Hz	0.08261	Magnetic_energy	0.0043	Pressure_energy_band_16_Hz
0.10975	Gyro_energy_band_14_Hz	0.08147	MelFilterBank#39	0.00428	Pressure_energy_band_10_Hz
				0.00425	Pressure_energy_band_13_Hz
				0.00403	Pressure_energy_band_17_Hz

TABLE 7.3: Résultats de la sélection par corrélation

Score	Descripteur	Score	Descripteur	Score	Descripteur
0.243	Pressure_mean	0.349	Magnetic_energy_band_16_Hz	0.337	Accel_energy_band_6_Hz
0.285	Gyro_variance	0.349	Pressure_energy_band_24Hz	0.337	Pressure_energy_band_18_Hz
0.313	MelFilterBank#1	0.349	Magnetic_energy_band_11_Hz	0.336	MelFilterBank#17
0.322	Accel_energy_band_3_Hz	0.349	MelFilterBank#10	0.336	Accel_energy_band_11_Hz
0.329	Magnetic_mean	0.349	Gyro_energy_band_9_Hz	0.336	Gyro_energy_band_15_Hz
0.335	MelFilterBank#6	0.348	Accel_energy_band_10_Hz	0.336	Pressure_energy_band_16_Hz
0.341	Gyro_energy_band_2_Hz	0.348	Magnetic_energy_band_20_Hz	0.335	Pressure_energy_band_13_Hz
0.344	Accel_variance	0.348	Accel_energy_band_22_Hz	0.335	Accel_energy_band_17_Hz
0.347	Magnetic_variance	0.348	MelFilterBank#12	0.335	MelFilterBank#32
0.35	MelFilterBank#0	0.347	Gyro_energy_band_7_Hz	0.334	Accel_energy_band_8_Hz
0.353	Accel_mean	0.347	Magnetic_energy_band_13_Hz	0.334	Pressure_energy_band_17_Hz
0.353	Gyro_energy_band_3_Hz	0.347	Magnetic_energy_band_5_Hz	0.334	Gyro_energy_band_18_Hz
0.354	MelFilterBank#2	0.347	Accel_energy_band_5_Hz	0.334	MelFilterBank#16
0.355	Accel_energy_band_23_Hz	0.346	MelFilterBank#39	0.333	Pressure_energy_band_8_Hz
0.355	Gyro_energy_band_1_Hz	0.346	Magnetic_energy_band_10_Hz	0.333	Accel_energy_band_15_Hz
0.355	MelFilterBank#9	0.346	Gyro_energy_band_11_Hz	0.333	Pressure_energy
0.355	Magnetic_energy_band_2_Hz	0.346	Accel_energy_band_2_Hz	0.332	Gyro_energy_band_19_Hz
0.356	Accel_energy_band_4_Hz	0.345	MelFilterBank#13	0.332	MelFilterBank#21
0.356	Magnetic_energy_band_24Hz	0.345	Magnetic_energy_band_18_Hz	0.332	Gyro_energy_band_25Hz
0.355	Gyro_energy_band_10_Hz	0.345	Accel_energy_band_19_Hz	0.331	Gyro_energy_band_16_Hz
0.355	MelFilterBank#3	0.345	Magnetic_energy_band_9_Hz	0.331	MelFilterBank#35
0.355	Accel_energy_band_21_Hz	0.344	Pressure_energy_band_22_Hz	0.331	Pressure_energy_band_3_Hz
0.355	Magnetic_energy_band_17_Hz	0.344	Accel_energy_band_7_Hz	0.33	Gyro_energy_band_20_Hz
0.354	Gyro_energy_band_4_Hz	0.344	Magnetic_energy_band_21_Hz	0.33	MelFilterBank#24
0.354	MelFilterBank#7	0.344	MelFilterBank#19	0.33	Gyro_energy_band_24Hz
0.354	Magnetic_energy_band_3_Hz	0.343	Gyro_energy_band_12_Hz	0.329	Gyro_energy_band_17_Hz
0.354	Accel_energy	0.343	Pressure_energy_band_2_Hz	0.329	MelFilterBank#18
0.354	Magnetic_energy_band_1_Hz	0.343	Magnetic_energy_band_15_Hz	0.329	Gyro_energy_band_21_Hz
0.353	MelFilterBank#4	0.343	Magnetic_energy_band_12_Hz	0.328	Pressure_energy_band_5_Hz
0.353	Gyro_energy	0.342	Accel_energy_band_14_Hz	0.328	MelFilterBank#31
0.353	Accel_energy_band_1_Hz	0.342	MelFilterBank#30	0.328	Gyro_energy_band_22_Hz
0.353	Magnetic_energy_band_4_Hz	0.342	Magnetic_energy_band_6_Hz	0.327	Pressure_energy_band_25Hz
0.353	MelFilterBank#11	0.342	Gyro_energy_band_13_Hz	0.327	Pressure_energy_band_11_Hz
0.353	Gyro_mean	0.341	Pressure_energy_band_20_Hz	0.326	MelFilterBank#37
0.353	Accel_energy_band_24Hz	0.341	Accel_energy_band_20_Hz	0.326	Pressure_energy_band_7_Hz
0.352	Magnetic_energy_band_23_Hz	0.341	MelFilterBank#15	0.326	Pressure_energy_band_15_Hz
0.352	Pressure_variance	0.341	Magnetic_energy	0.325	Pressure_energy_band_10_Hz
0.352	Magnetic_energy_band_14_Hz	0.34	Pressure_energy_band_1_Hz	0.325	Pressure_energy_band_9_Hz
0.352	MelFilterBank#8	0.34	Accel_energy_band_12_Hz	0.324	MelFilterBank#22
0.352	Gyro_energy_band_5_Hz	0.34	Magnetic_energy_band_22_Hz	0.324	Pressure_energy_band_4_Hz
0.352	Accel_energy_band_18_Hz	0.34	Gyro_energy_band_14_Hz	0.324	Pressure_energy_band_21_Hz
0.351	Magnetic_energy_band_7_Hz	0.339	MelFilterBank#36	0.323	Pressure_energy_band_14_Hz
0.351	MelFilterBank#5	0.339	Magnetic_energy_band_19_Hz	0.323	MelFilterBank#34
0.351	Gyro_energy_band_8_Hz	0.339	Accel_energy_band_9_Hz	0.322	Pressure_energy_band_19_Hz
0.351	Magnetic_energy_band_25Hz	0.339	Pressure_energy_band_23_Hz	0.322	MelFilterBank#28
0.351	Accel_energy_band_13_Hz	0.338	Pressure_energy_band_12_Hz	0.321	MelFilterBank#23
0.35	Magnetic_energy_band_8_Hz	0.338	Accel_energy_band_16_Hz	0.321	MelFilterBank#38
0.35	MelFilterBank#14	0.338	MelFilterBank#20	0.32	MelFilterBank#33
0.35	Gyro_energy_band_6_Hz	0.338	Gyro_energy_band_23_Hz	0.319	MelFilterBank#25
0.35	Accel_energy_band_25Hz	0.337	Pressure_energy_band_6_Hz	0.318	MelFilterBank#29
				0.318	MelFilterBank#26
				0.317	MelFilterBank#27

Fiche préalable de Traitement* v2.6

Correspondant informatique et libertés mutualisé
des établissements universitaires du PRES Université de Grenoble

Utilisez la fiche annotée pour renseigner cette fiche préalable de traitement

(*) Information aux personnes concernées par les formalités d'un traitement (responsable du traitement, personnes en charge ou opérateurs de la mise en œuvre, référents informatique et libertés, destinataires éventuels des données).
L'instruction des formalités de déclaration des traitements mis en œuvre par les établissements constitue un traitement déclaré et porté au registre du Correspondant informatique et libertés (CIL) mutualisé des établissements.
Certaines de vos données à caractère personnel peuvent apparaître dans le présent document et dans le registre du CIL, accessible au public.
Conformément à la loi « informatique et libertés » du 6 janvier 1978 modifiée, vous pouvez exercer vos droit d'accès et de rectification, pour les données qui vous concernent, auprès du CIL par mail à CIL@grenet.fr (Sujet/Objet : « CIL : DEMANDE D'EXERCICE DES DROITS »)

ENTETE DE FICHE**

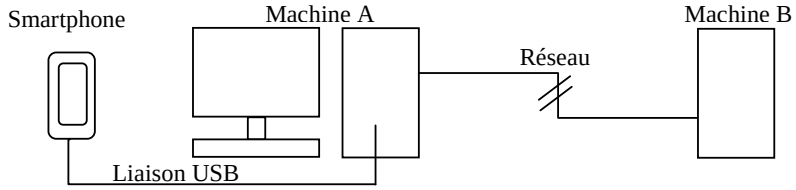
DATE DEMANDE	ETABLISSEMENT	SERVICE OU LABORATOIRE	NOM DU TRAITEMENT	
21/01/2013	*UJF*	*LIG-GETALP*	*Collecte et traitement de données multimodales sur smartphone*	
VERSION	**DATE**	**RELAISCIL/CIL**	**RESP. FONCTIONNEL (MOE)**	**VALIDATION / DATE ****
Version 1	*21/01/2013*	*Bernard MARTINET/ Patrick GUILHOT*	*Laurent BESACIER (resp.) Stéphan TASSART, ST (coresponsable.)*	*21/01/2013*

(**) L'entête de la fiche est renseigné par le demandeur. Cette fiche doit être préalablement validée par le responsable de la mise en œuvre du traitement pour que le traitement puisse être déclaré (porté au registre du CIL) : fiche datée et signée ou réf. courrier ou mail de validation du responsable du projet

CORPS DE FICHE

TYPE FICHE DE TRAITEMENT	☐ Création ☐ Modification / Mise à jour ☐ Suppression	*N°déclaration (réservé CIL)* réf. du traitement modifié : réf. traitement supprimé :	
FORMALITE CNIL	☐ Dispense ☐ Norme simplifiée ☐ ☐ Autorisation unique ☐ ☐ Acte réglementaire unique ☐	DI-☐ ☐ ☐ NS-☐ ☐ AU-☐ ☐ RU-☐ ☐	*Réservé CIL*
	☐ Normale ☐ Avis CNIL ☐ Autorisation CNIL	DN DAv DAu	*Réservé CIL*
RESPONSABLE DE TRAITEMENT	Laurent BESACIER		
CARACTERE OBLIGATOIRE	Non		
FINALITE(S)	La collecte de données s'inscrit dans un projet de thèse CIFRE dont le sujet est « Embedded Multimodal Scene Recognition » (reconnaissance de scènes multimodale embarquée). Il s'agit de détecter et d'identifier les changements de scènes environnant l'utilisateur et son smartphone à partir des données mesurées par les capteurs du smartphone. La finalité de cette reconnaissance est de proposer à l'utilisateur des services adaptés à l'environnement qui l'entoure. Un exemple est l'adaptation du volume de sonnerie du téléphone en fonction du type d'environnement (par exemple : lors d'une réunion au travail). Une scène est définie comme la combinaison de la localisation géographique		

	de l'utilisateur, de son activité et de son contexte social. Les scènes seront « mesurées » par des capteurs présents sur le téléphone et par certains éléments de l'activité du téléphone (les détails des données collectées sont présentées dans la section « Données traitées »). **La collecte des données servira à l'apprentissage des modèles des différentes scènes. Une fois que les modèles auront appris à partir d'un sous-ensemble des données collectés, ils pourront être testés sur un autre sous-ensemble pour réaliser la détection et l'identification des différentes scènes.**
DETAILS DU TRAITEMENT	**Le traitement de la présente déclaration consiste à :** • **réaliser une collecte de données à partir d'un smartphone de manière continue pendant une période donnée ; cette collecte portera sur les capteurs du smartphone, des éléments d'activité du smartphone et les annotations réalisées par les sujets pendant l'expérimentation,** • **rendre anonymes les données sensibles qui le nécessitent et ne garder que ces données anonymes et des données non sensibles,** • **transférer les données vers une ou plusieurs machines pour les y stocker,** • **effectuer le travail de recherche sur les données stockées.** **La première partie du traitement consiste en l'acquisition des données. Celle-ci sera faite sur un smartphone fourni à l'utilisateur avec une application pour l'enregistrement des données développée dans le cadre de ce projet. Cette application permettra à l'utilisateur de choisir les types des données à enregistrer (par exemple audio, accélérations, pression ambiante) puis de lancer l'enregistrement. Une fois l'enregistrement lancé, l'utilisateur pourra quitter l'application et utiliser son smartphone normalement. A tout moment, l'utilisateur sera en mesure de visualiser l'état de l'enregistrement (en cours ou non) ainsi que les types de données en cours d'enregistrement. L'utilisateur pourra également stopper l'enregistrement soit partiellement, en n'arrêtant l'enregistrement que de certains types de données, soit en coupant complètement l'enregistrement. Dans les deux cas, l'utilisateur pourra relancer l'enregistrement des types de données arrêtés.** **En plus des données enregistrées automatiquement, l'utilisateur sera sollicité pour renseigner des annotations sur la scène courante, au moment où la scène change. Cette information est importante car elle servira de référence pour les données enregistrées dans les différentes scènes. Cependant, l'action d'annoter est complètement volontaire et il y a donc le risque que l'utilisateur ne veuille pas, ne puisse pas ou oublie simplement d'annoter la scène. Dans le cas où l'utilisateur souhaiterait annoter la scène courante, il pourra le faire par le biais de l'application d'enregistrement qui proposera une interface pour réaliser cette annotation. L'utilisateur sera guidé dans les informations à fournir. L'ensemble des éléments à annoter est renseigné dans la section « Données traitées ».** **Une fois l'enregistrement lancé, et sous réserve que l'utilisateur ne le stoppe pas ou que le système d'exploitation ne l'arrête pas, l'enregistrement des données sera continu. Cette continuité permettra d'avoir de longues séquences avec des changements de scènes et donc d'avoir des données réelles et ainsi de réduire le biais expérimental de celles-ci. Cependant, il n'est pas exclu de mettre en place des scénarios de scènes à suivre (scènes à « réaliser » suivant des indications temporelles) afin de garantir un volume minimum de données pour certaines scènes et un minimum de confiance dans les annotations de ces scènes.** **La phase d'acquisition durera aussi longtemps que la durée prévue dans la section « Durée du traitement ». Les utilisateurs impliqués dans cette campagne ne seront que les expérimentateurs, soit le doctorant et ses trois encadrants de thèse. Ces utilisateurs sont directement liés au traitement et sont donc au courant de l'ensemble des étapes de celui-ci.** **La seconde étape du traitement consiste à rendre anonymes les données que nous appelons sensibles. Toutes les données collectées ne sont pas sensibles. Ainsi, dans le cadre du présent traitement, les données sensibles sont le signal audio, les données de localisation ainsi que des chaînes de caractère telles que**

	des numéros de téléphone, des identifiants d'appareils, ou encore les noms d'applications utilisées. La liste des données enregistrées à anonymiser est présentée dans la section « Données traitées ». Bien que cette étape soit présentée comme la seconde étape, elle sera réalisée (pour l'essentiel) au moment de l'acquisition, sur le smartphone. Différentes techniques d'anonymisation sont envisagées en fonction des types de données. Pour les chaînes de caractères ou d'entiers tels que les noms, numéros de téléphone et identifiants divers, l'anonymisation se fera par application d'une fonction de hachage. Concernant le signal audio, celui-ci ne sera pas enregistré au format brut ; des paramètres seront calculés en appliquant au signal en temps-réel des transformations fréquentielles irréversibles. Ces paramètres serviront au travail de recherche qui suivra et seuls ces paramètres seront sauvegardés au moment de l'acquisition. Enfin, concernant les données de localisation, une première transformation continue réversible sera appliquée à ces données pour les « encrypter » le temps du stockage sur le smartphone. Puis, lors du transfert des données sur machine, une nouvelle transformation continue irréversible sera appliquée. Les détails sur toutes les mesures d'anonymisation sont présentés dans la section « Sécurité des données ». **La troisième étape du traitement dans la liste consiste à transférer les données du smartphone vers une ou plusieurs machines. Le schéma ci-dessous illustre le transfert. En pratique, ce transfert se fera par liaison USB du smartphone vers la machine A qui servira de relais pour le transfert des données vers la machine B. La machine B sera dédiée au stockage des données et sera accédée depuis la machine A via une connexion réseau. La machine A « extraira » des labels génériques tels que « ville », « campagne », « intérieur », « extérieur » à partir des coordonnées GPS. En outre, elle réalisera l'anonymisation des coordonnées GPS par application d'une transformation continue irréversible. Les détails sur les procédés d'anonymisation et de sécurisation des différents transferts entre machines seront détaillés dans la section « Sécurité des données ». Le transfert sera réalisé quotidiennement afin de limiter la durée de stockage sur le smartphone.**  **Enfin, la dernière étape représente le travail de recherche réalisé sur ces données. Celui-ci consistera à rechercher des propriétés caractéristiques des scènes à partir des données. Ces propriétés seront représentées par des paramètres qui seront calculés et des modèles qui seront établis à partir de ces paramètres. Ces paramètres et modèles sont l'objet du travail de recherche et ne sont donc pas encore définis. Egalement, des algorithmes seront créés pour réaliser la détection et l'identification des scènes.**
LIENS AVEC D'AUTRES TRAITEMENTS	**A ce jour, les données enregistrées ne seront exploitées que dans le cadre du présent traitement. Cependant, ces données et leur protocole d'acquisition, seront vraisemblablement une plus-value de la thèse. Ainsi, nous souhaiterions pouvoir utiliser les données dans le cadre d'autres traitements, notamment en les rendant publiques afin que d'autres laboratoires puissent les utiliser. Le cas échéant, les données transmises ne seraient que des données anonymes.**
DATE SOUHAITEE DE MISE EN ŒUVRE	**Février 2013**
RECURRENCE DU TRAITEMENT	**Le traitement aura lieu une première fois avec un ensemble d'utilisateurs réduit aux personnes impliquées dans sa conception (appelés les expérimentateurs) et composé du doctorant et de ses encadrants de thèse.** **Nous envisageons de réaliser ce traitement une seconde fois, en faisant cette fois-ci appel à des volontaires. Ce second traitement donnera lieu à une mise à jour de la présente fiche de traitement pour tous les aspects qui le**

<table>
<tr><td></td><td colspan="4">nécessiteront.</td></tr>
<tr><td>DUREE DU TRAITEMENT OU DATE DE FIN OU DATE DE MISE A JOUR (SI EVOLUTION PREVUE)</td><td colspan="4">Différentes durées s’appliquent :
la phase d’acquisition des données durera 3 mois,
la phase de traitement des données commencera en même temps que la phase d’acquisition et durera jusqu’à la fin de la thèse, soit début mars 2015,
la phase de conservation des données anonymisées ou qui ne nécessitent pas d’anonymisation durera aussi longtemps que la phase de traitement.
Par ailleurs, comme indiqué dans la section précédente, une seconde collecte de données est envisagée, ainsi une mise à jour sera vraisemblablement faite à cette fiche au cours du deuxième trimestre de l’année 2013. Cette date est purement indicative et pourra évoluer en fonction de l’évolution du projet et des résultats de la présente campagne.</td></tr>
<tr><td>CATEGORIE DES PERSONNES CONCERNEES</td><td colspan="4">Il s’agit des expérimentateurs de ce traitement, soit le doctorant et ses 3 encadrants de thèse.</td></tr>
<tr><td>INFORMATION AUX PERSONNES CONCERNEES</td><td colspan="4">Dans le cadre du présent traitement, les utilisateurs seront les expérimentateurs et sont donc au courant des différents éléments du traitement par le fait qu’ils participent à l’élaboration du protocole du présent traitement.</td></tr>
<tr><td>SERVICE OU PERSONNES AUPRES DUQUEL S’EXERCE LE DROIT D’ACCES</td><td colspan="4">Les données collectées au cours de ce traitement seront anonymisées lorsque la donnée originelle est considérée comme sensible. Ces données anonymes, ainsi que celles qui ne sont pas sensibles et qui ne seront pas anonymisées, seront sauvegardées sur la machine B qui sera accessible aux expérimentateurs.</td></tr>
<tr><td rowspan="8">DONNEES TRAITEES</td><td></td><td></td><td></td><td></td></tr>
<tr><td>DONNEES OU CATEGORIES</td><td>ORIGINE / SOURCE*</td><td>DUREE DE CONSERVATION**</td><td>DESTINATAIRES***</td></tr>
<tr><td>Accélérations selon 3 axes orthogonaux toutes les 20 ms</td><td>Présent traitement</td><td>Jusqu’à fin de thèse</td><td>LIG, STMicroelectronics</td></tr>
<tr><td>Champs magnétiques selon 3 axes orthogonaux toutes les 20 ms</td><td>Présent traitement</td><td>Jusqu’à fin de thèse</td><td>LIG, STMicroelectronics</td></tr>
<tr><td>Vitesses de rotation selon 3 axes orthogonaux toutes les 20 ms</td><td>Présent traitement</td><td>Jusqu’à fin de thèse</td><td>LIG, STMicroelectronics</td></tr>
<tr><td>Donnée de luminosité ambiante</td><td>Présent traitement</td><td>Jusqu’à fin de thèse</td><td>LIG, STMicroelectronics</td></tr>
<tr><td>Donnée de proximité du smartphone à un objet (champ de l’ordre de quelques cm)</td><td>Présent traitement</td><td>Jusqu’à fin de thèse</td><td>LIG, STMicroelectronics</td></tr>
<tr><td>Pression ambiante</td><td>Présent traitement</td><td>Jusqu’à fin de thèse</td><td>LIG, STMicroelectronics</td></tr>
<tr><td>Paramètres (signal audio brut ayant subi une transformation fréquentielle irréversible)
Anonymisation : calcul et stockage des paramètres ; le</td><td>Présent traitement</td><td>Jusqu’à fin de thèse</td><td>LIG, STMicroelectronics</td></tr>
</table>

	signal brut n'est pas conservé			
	Casque audio : événement de branchement ou débranchement	Présent traitement	Jusqu'à fin de thèse	LIG, STMicroelectronics
	Batterie : état faible/bon, chargeur connecté ou non	Présent traitement	Jusqu'à fin de thèse	LIG, STMicroelectronics
	Ecran smartphone verrouillé : oui/non	Présent traitement	Jusqu'à fin de thèse	LIG, STMicroelectronics
	Application utilisée en premier plan : nom de l'application Anonymisation du nom par application d'une fonction de hachage	Présent traitement	Jusqu'à fin de thèse	LIG, STMicroelectronics
	Appels : état (entrant, en cours, pas d'appel), numéro entrant, état du service (volontairement coupé, pas de service, urgences seulement, service normal) Anonymisation du numéro par application d'une fonction de hachage	Présent traitement	Jusqu'à fin de thèse	LIG, STMicroelectronics
	SMS reçus : numéro de l'émetteur Anonymisation du numéro par application d'une fonction de hachage	Présent traitement	Jusqu'à fin de thèse	LIG, STMicroelectronics
	Bluetooth : état (actif/inactif), appareils environnants détectés (nom et adresses MAC) Anonymisation : noms et adresses MAC par application d'une fonction de hachage	Présent traitement	Jusqu'à fin de thèse	LIG, STMicroelectronics
	Wifi : état (actif/inactif), état connexion réseau, points d'accès alentours (identifiants) Anonymisation : identifiants par application d'une fonction de hachage	Présent traitement	Jusqu'à fin de thèse	LIG, STMicroelectronics
	Data : état connexion, type connexion (mobile/wifi), sens transfert (entrant et/ou sortant, « dormant »)	Présent traitement	Jusqu'à fin de thèse	LIG, STMicroelectronics
	Localisation par GPS/Wifi : état (actif/inactif), données localisation (latitude, longitude, altitude, vitesse, précision, horodatage), labels génériques associés aux coordonnées (ex : ville, campagne, intérieur, extérieur) Anonymisation : transformation continue irréversible des coordonnées	Présent traitement	Jusqu'à fin de thèse	LIG, STMicroelectronics

	Localisation par antennes téléphoniques : identifiant antenne connectée, code de localisation de l'antenne Anonymisation : identifiant antenne et code de localisation, par application d'une fonction de hachage	Présent traitement	Jusqu'à fin de thèse	LIG, STMicroelectronics
	Annotations : • localisation/mouvement : liste proposée + texte libre • activité de l'utilisateur : liste proposée + texte libre • contexte social : seul ou non, interaction avec des personnes ou non, personnes connues ou non • scène habituelle ou non ; fréquence de la scène	Présent traitement	Jusqu'à fin de thèse	LIG, STMicroelectronics

SECURITE DES DONNEES	**La sécurité des données doit être réalisée à chaque étape du traitement, à savoir :** • **au moment de l'acquisition et lors du stockage des données sur le téléphone,** • **lors du transfert des données du téléphone vers la machine A,** • **lors du passage temporaire des données sur la machine A,** • **lors du transfert des données de la machine A vers la machine B,** • **lors du stockage des données sur la machine B,** • **lors du traitement ultérieur des données.** **Pendant l'acquisition des données et tant que les données sont stockées sur le téléphone, le risque provient essentiellement de la perte du téléphone. Plusieurs solutions sont mises en place. La première solution mise en place consiste en l'anonymisation des données les plus sensibles réalisée directement sur le smartphone. Ainsi, pour les données qui peuvent être assimilées à des chaînes de caractères, une fonction de hachage est implémentée. L'algorithme utilisé est le Message Digest 5 (MD5). L'anonymisation par application d'une fonction de hachage consiste à appliquer une transformation irréversible à la donnée à anonymiser et à ne garder que le résultat de cette transformation. Par le caractère irréversible de la transformation, il est impossible de retrouver la donnée initiale à partir de la donnée transformée. Ensuite, concernant le signal audio, l'anonymisation sera réalisée par calcul de paramètres. Ces paramètres seront calculés en appliquant sur des portions de signal comprises entre 200 et 500 ms des transformations fréquentielles irréversibles. Par ailleurs, les utilisateurs auront la possibilité de couper l'enregistrement des données audio quand ils le souhaiteront. Il est également prévu de couper automatiquement l'enregistrement lors d'un appel. Enfin, une transformation continue réversible sera appliquée aux coordonnées GPS pendant le stockage des coordonnées sur le smartphone en vue de les encrypter. Outre ces mesures d'anonymisatrion, la destruction automatique et régulière des fichiers de données sur le smartphone sera réalisée. Les conditions de destruction seront soit après le transfert automatique des données vers la machine A si celui-ci est effectué, soit à une heure fixée si le transfert n'a pas eu lieu à ce moment-là. L'utilisateur pourrait utiliser cette fonctionnalité de manière dérivée pour détruire des données qu'il ne souhaite pas transférer, simplement en ne réalisant pas le transfert dans la plage de temps envisagée et en laissant donc l'application détruire les données.** **Le transfert des données du smartphone vers la machine A se fera par liaison USB. La machine A sera située dans un bureau fermé à clef et dont la clef n'est accessible que par un nombre restreint de personnes (par ailleurs, le bureau se situe dans un des bâtiments du LIG dont l'accès est contrôlé par badge). La machine A sera protégée par une authentification par mot de passe. Lors de la connexion du smartphone à la machine A, un script automatique sera chargé de récupérer les données du smartphone, les copier temporairement sur la machine A, les « décrypter » en appliquant la transformation inverse à celle qui avait été appliquée sur le smartphone, associer des labels génériques aux coordonnées (voir la section « Données traitées » pour plus de détails), appliquer une transformation continue irréversible aux coordonnées afin de les rendre anonymes, transférer ces coordonnées anonymes sur la machine B et, finalement, détruire les données du téléphone et de la machine A.** **Le transfert des données de la machine A vers la machine B se fera par un accès réseau sécurisé. La machine B sera entreposée dans un local à accès sécurisé dans un des bâtiments du LIG. Cette machine sera protégée par une authentification par mot de passe.** **Ainsi, lors du traitement ultérieur, les données utilisées seront celles stockées sur la machine B. Elles seront transférées par liaison réseau sécurisée vers une machine pour faire le traitement de recherche (machine qui sera elle aussi protégée par mot de passe, située dans une pièce fermée à clef).**

CONFIDENTIALITE	**La confidentialité est nécessaire pour ce traitement dans la mesure où certaines données collectées sont sensibles. Comme nous l'avons déjà présenté dans les sections précédentes qui détaillent le traitement, des procédés d'anonymisation sont mis en place dès l'acquisition des données afin de ne sauvegarder que des données non sensibles ou anonymisées.** **En outre, la confidentialité des données doit également être assurée lors du transfert des données du smartphone vers la machine B en passant par la machine A. Pour cela, la mise en place d'un script automatique qui réalise ce transfert devrait permettre d'éviter toute manipulation des données au moment du transfert et ainsi garantir la confidentialité de celles-ci. Ce script sera approuvé par l'ensemble des expérimentateurs.**
CATEGORIES DES PERSONNES OU SERVICES CHARGES DE LA MISE EN ŒUVRE (MOE)	**La MOE sera réalisée par l'équipe GETALP du LIG via M. Laurent BESACIER ainsi que par la société STMicroelectronics via M. Stéphan TASSART.**
RESPONSABLE DE LA MISE EN ŒUVRE	**Laurent BESACIER, Stéphan TASSART**
MOYENS DE MISE EN ŒUVRE	**Dans un premier temps, le traitement nécessite l'utilisation de smartphones équipés de l'application d'enregistrement des données. Les smartphones sont fournis par l'UJF et l'application est développée par le doctorant et sera installée sur les smartphones.** **Concernant le transfert des données sur la machine A, le câble qui réalise la liaison et ladite machine A sont nécessaires. Les câbles seront fournis avec les smartphones tandis que la machine A est mise à disposition dans les locaux du LIG. Sur cette machine, un compte protégé par mot de passe sera créé pour y transférer les données. Le script du transfert automatique des données sera réalisé par les expérimentateurs.** **Enfin, la machine B nécessaire au stockage sera entreposée au LIG et accessible par une connexion réseau.**
SOUS-TRAITANCE	**Pas de sous-traitance.**
TRANSFERT DES DONNEES HORS UNION EUROPEENNE	- **Comme indiqué dans la section « Liens avec d'autres traitements », nous souhaitons pouvoir utiliser ces données et en particulier les rendre publiques afin que, par exemple, d'autres laboratoires puissent les utiliser dans leurs propres expérimentations ; cependant, rien n'est défini pour le moment ;** - **Le cas échéant, les données qui seraient transférées sont celles stockées sur la machine B, à savoir d'une part, des données brutes qui n'ont pas besoin d'être anonymisées et, d'autre part, des données anonymisées ; AUCUNE DONNEE SENSIBLE NE SERA TRANSFEREE SANS AVOIR ETE ANONYMISEE ;**

Remarques Observation Informations complémentaires	**Les informations ci-dessous ne figureront pas dans la déclaration du traitement et sont simplement un résumé de la réunion avec le CIL et relais CIL.** • **La ligne directrice du choix des types de données à enregistrer est « ni trop peu, ni pas assez » ; il ne faut prendre que ceux qui sont nécessaires au traitement ; ainsi on diminue le risque de perte des données ; dans notre cas, il est difficile de savoir avant cette première collecte quels types de données seront pertinents** • **Concernant les annotations lors de la seconde campagne, il n'est pas envisageable de laisser le choix à l'utilisateur de renseigner un texte libre car il y aurait un risque d'avoir des informations privées, donc sensibles.** • **Le problème de l'enregistrement du numéro des appels entrants a été soulevé. Plus généralement, cette question fait référence aux types de données enregistrés. Il a été suggéré de ne collecter que les données nécessaires et suffisantes (ni trop peu, ni trop) au traitement. Cependant, puisque notre projet est exploratoire, il est difficile de déterminer à l'avance quelles données seront pertinentes, c'est pourquoi nous préférons élargir au maximum le spectre des données enregistrées.** • **Lors de la seconde campagne, le transfert des données depuis les smartphones devra être revu. En particulier, l'utilisation d'une machine A unique paraît difficilement envisageable.** • **Les informations à fournir aux volontaires de la seconde campagne devront être fournies au CIL également. En particulier, préciser que lorsque les données sont anonymisées, il n'est plus possible de faire valoir son droit d'accès aux données.**

URecord

An Application for the Collection of Everyday Smartphone Data

David BLACHON

Presented by Stéphan TASSART

life.augmented

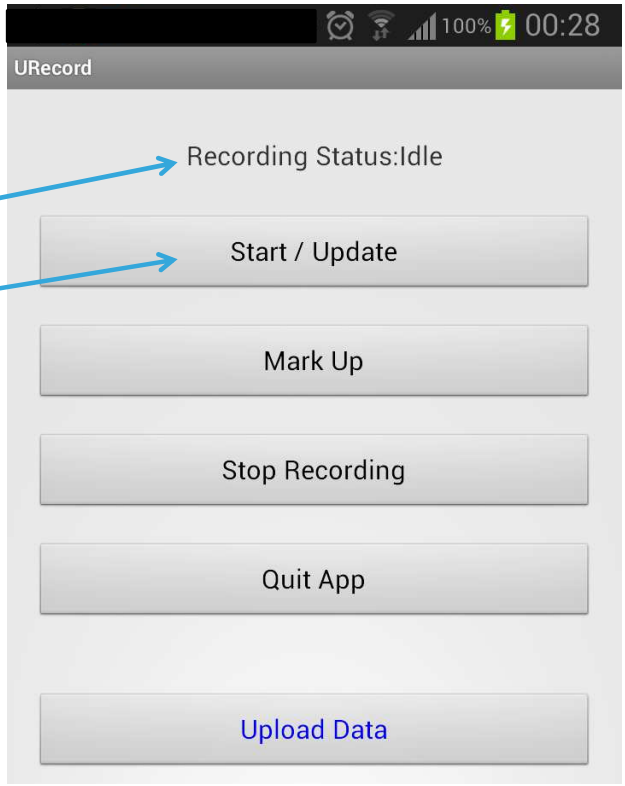
Step #1: Install the App

2

- From a website:
  - Visit membres-liglab.imag.fr/blachon/download/urecord.apk
  - Your smartphone should download the apk file and install it
    - By default, Android may block app installation not coming from Play Store, yet a dialog box should appear to invite you to accept the installation from different platforms; you should accept if you want to install the app from the website
- From an email:
  - We can send the application enclosed to an email that you will receive on your smartphone
  - Then Android should automatically install the app
- Once installed, you should see:

Step #2: Start a Recording 3

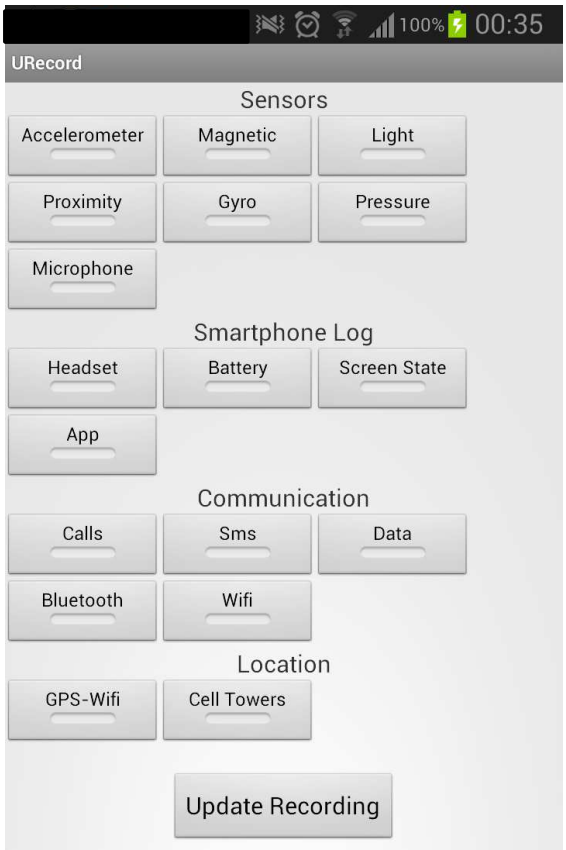
- Launch the application
- At this time, no recording yet
  - Recording Status informs at any time of the current status
- Start a Recording



life.augmented

Step #3: Select Sensors 4

- All available sensors appear
  - The more you record, the better
- When you select sensors, they change color

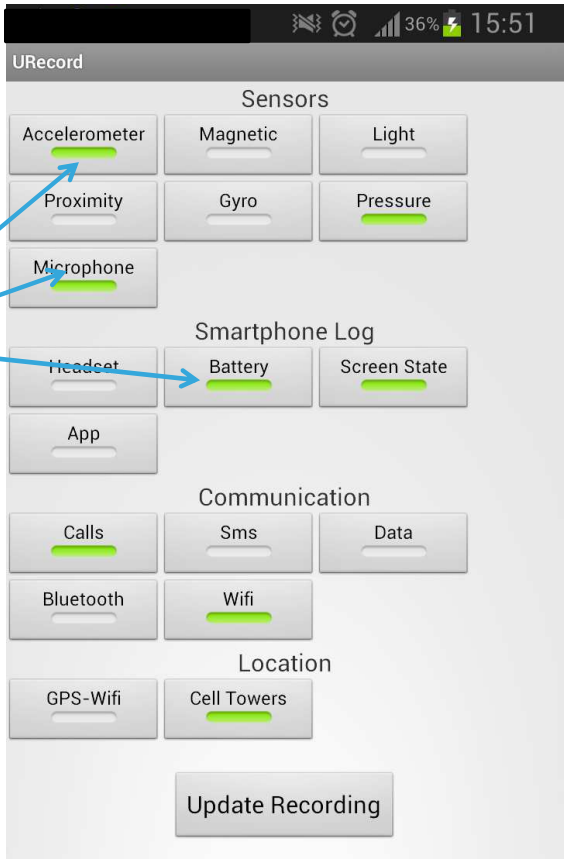


life.augmented

Step #3: Select Sensors

5

- All available sensors appear
  - The more you record, the better
  - For GPS and wifi, do not forget to turn them on before, the app can't turn them on by itself
- When you select sensors, they change color

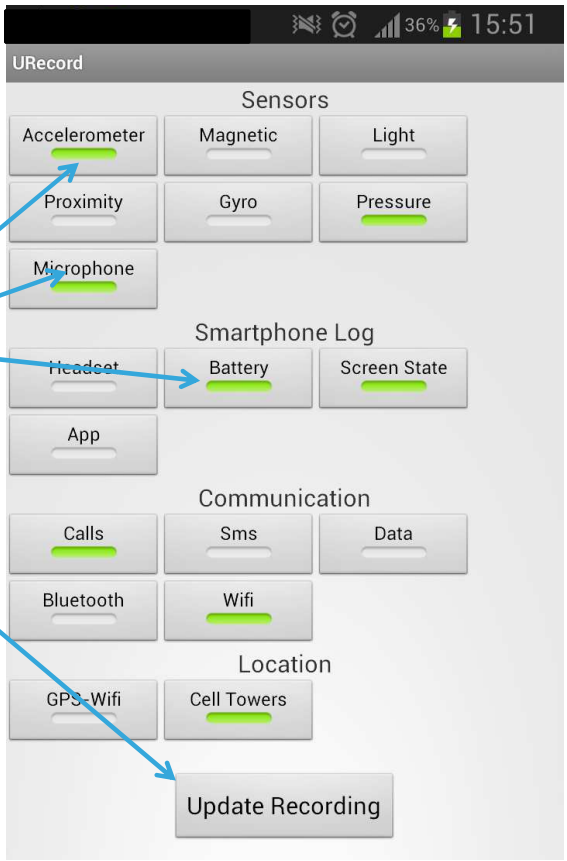


life.augmented

Step #3: Select Sensors

6

- All available sensors appear
  - The more you record, the better
  - For GPS and wifi, do not forget to turn them on before, the app can't turn them on by itself
- When you select sensors, they change color
- Then validate the selection for starting the record
  - You will be redirected to Home view

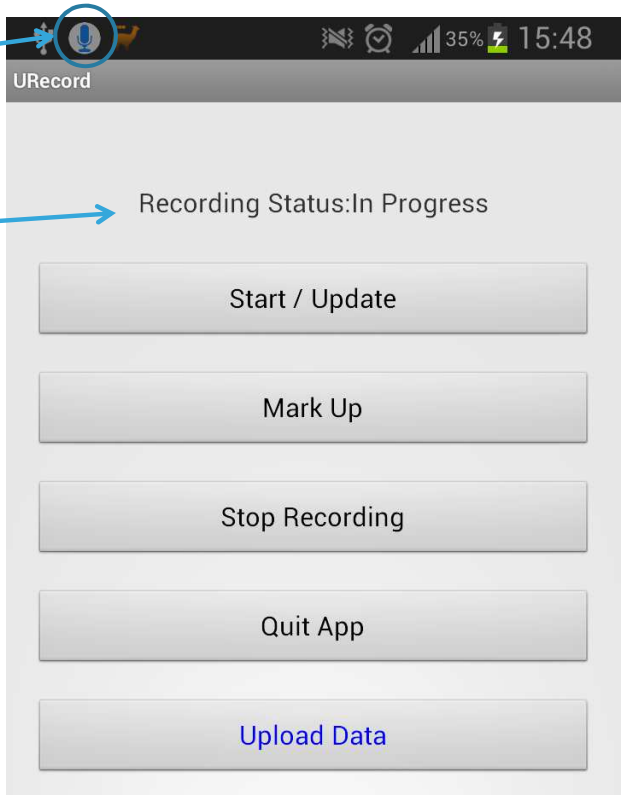


life.augmented

How do you know that a recording has been started?

7

- A notification appeared in the notification bar
- The Recording status now indicates: “In Progress”

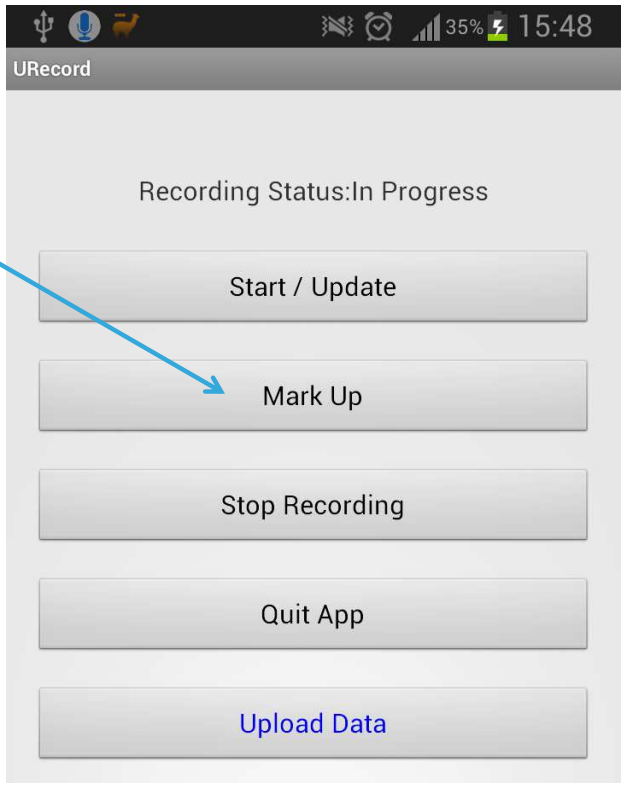


life.augmented

Step #4: Annotate the Current Scene

8

- Click the Mark Up button from Home view

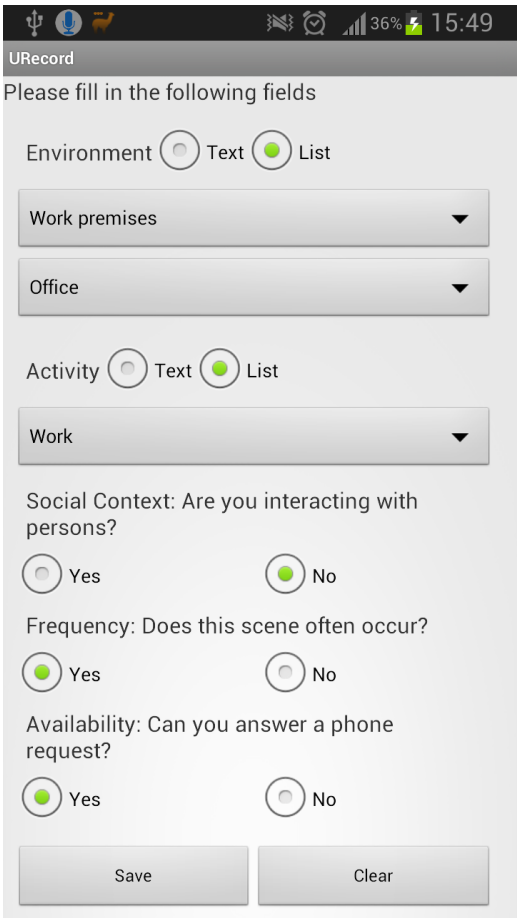


life.augmented

Step #4: Annotate the Current Scene

9

- Click the Mark Up button from Home view… then the Annotate view appears

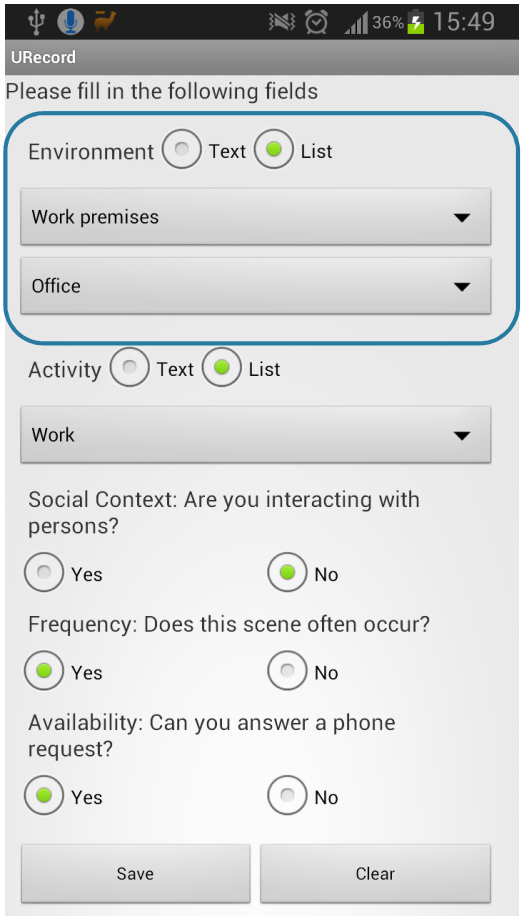


life.augmented

Step #4: Annotate the Current Scene

10

- Click the Mark Up button from Home view… then the Annotate view appears
- First, fill in the “Environment” section
  - Free Text Area (not recommended): Type labels of environment
  - List: Environment labels are grouped; first pick the group label then select the environment
  - E.g. “Office” is an environment from the “Work premises” group

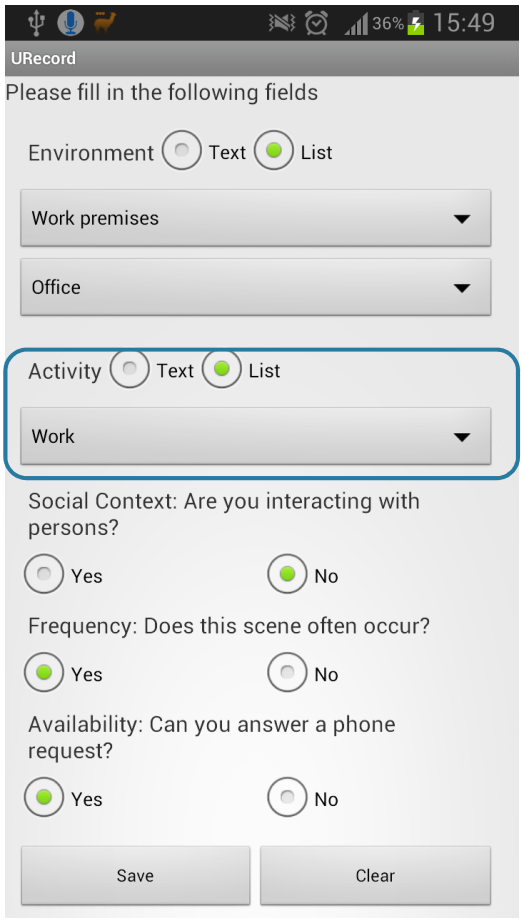


life.augmented

Step #4: Annotate the Current Scene

11

- Click the Mark Up button from Home view… then the Annotate view appears
- First, fill in the “Environment” section
- Secondly, fill in the Activity section
  - Free Text area (not recommended): type an activity label
  - List: pick an activity label
  - Activity labels are dynamically adapted loaded according to the environment group you picked;
  - E.g. because i picked “Work premises”, one of the activities is “Work”

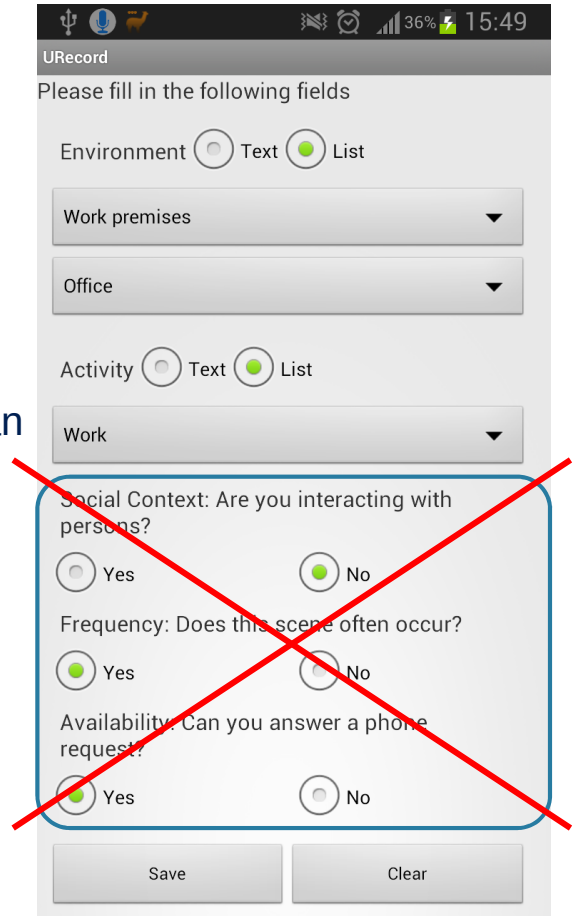


life.augmented

Step #4: Annotate the Current Scene

12

- Click the Mark Up button from Home view… then the Annotate view appears
- First, fill in the “Environment” section
- Secondly, fill in the Activity section
- Do not fill the yes/no questions: this is an experimental work and we don’t expect you to annotate this
- Save your annotation: you’ll be redirected to Home view
- Clear will reset this view

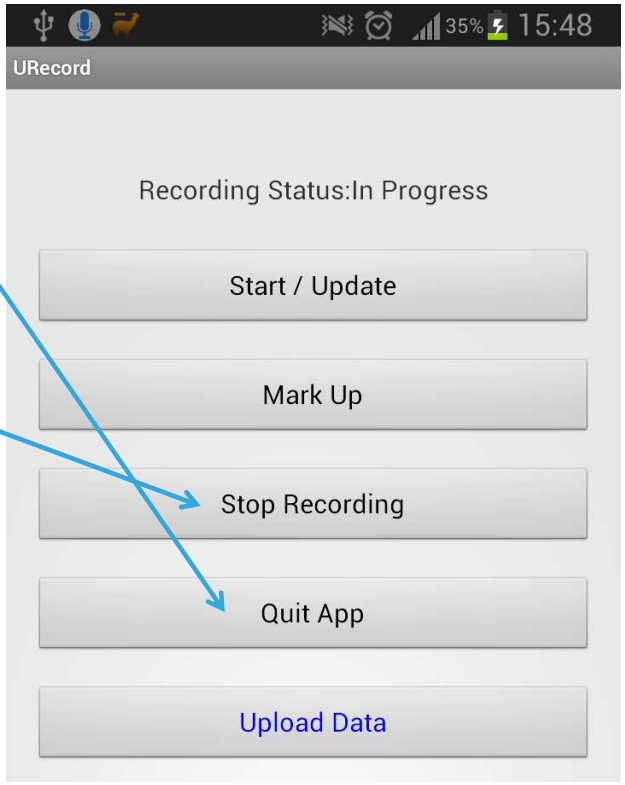


life.augmented

Step #5: Use your phone as usual

13

- Clicking “Quit” will allow you to keep recording while using other apps
  - The recording is a background task
  - Then, if you want to return to Urecord, you can click the notification that will redirect you to Home view
- Clicking “Stop” will stop the recording
  - The Recording Status will then display Idle
  - The notification will disappear from the notification bar

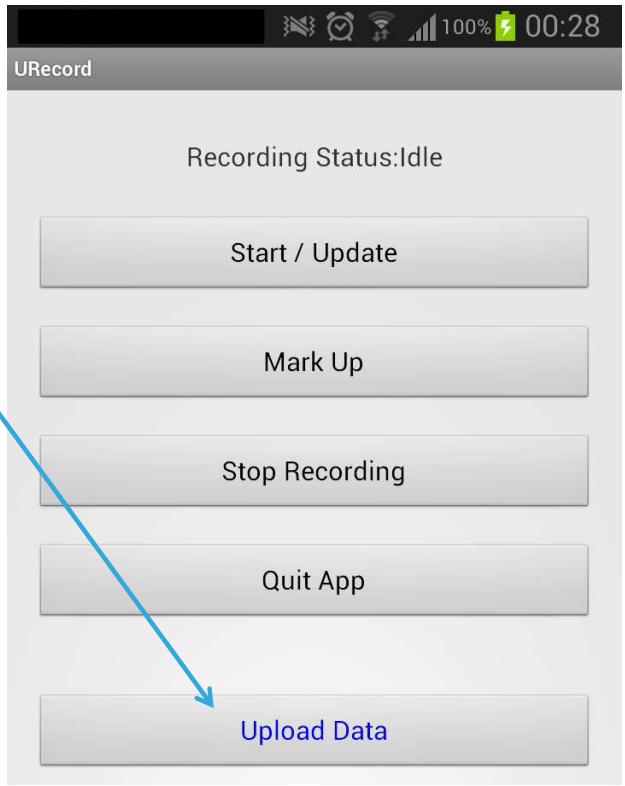


life.augmented

Step #6: Upload data

14

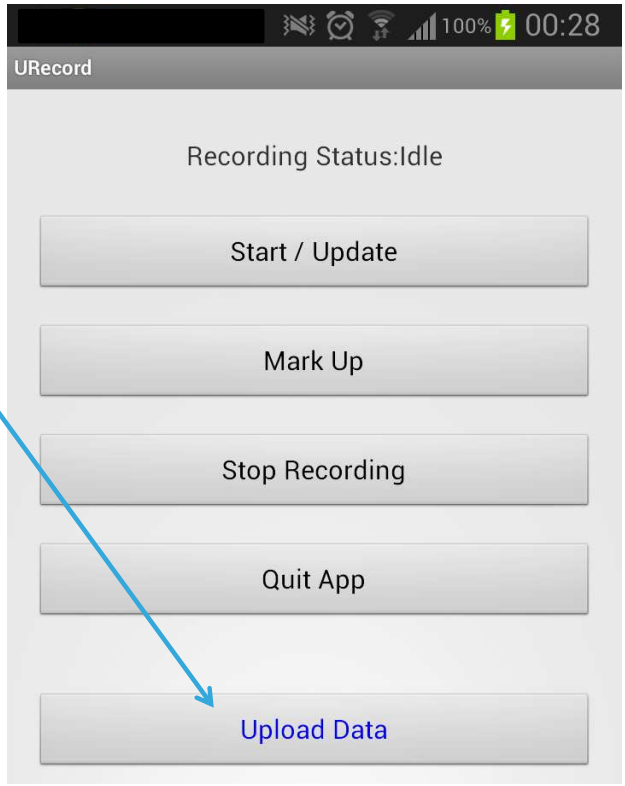
- Once you’ve finished a recording, you can upload data to our server
- From Home view, click Upload



life.augmented

Step #6: Upload data 15

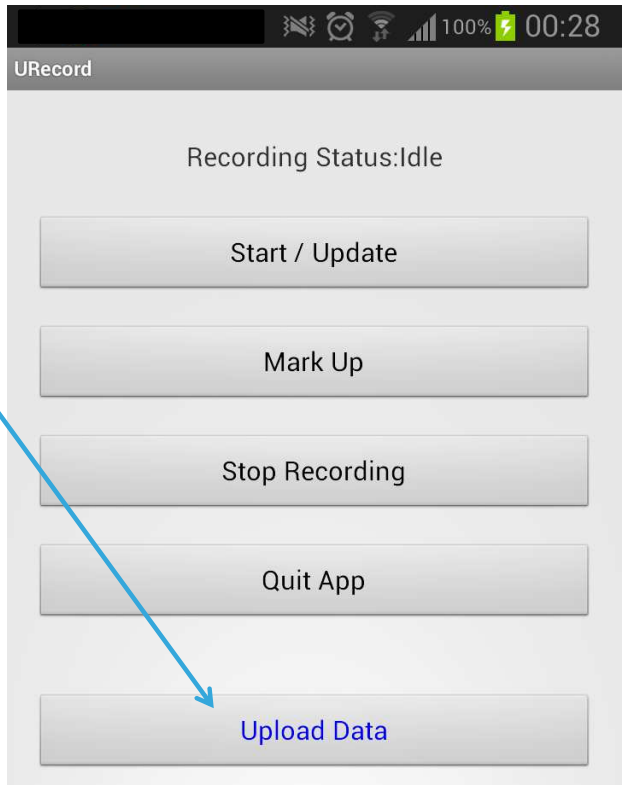
- Once you've finished a recording, you can upload data to our server
- From Home view, click Upload
- Upload guideline
  - Upload uses Wifi to transfer data, turn it on before upload
  - Estimated upload duration may vary according to your Wifi connection and amount of collected data, roughly it can take 1 to 2 hours
  - We recommend to upload data at night, when back home, and battery loading



life.augmented

Step #6: Upload data 16

- Once you've finished a recording, you can upload data to our server
- From Home view, click Upload
- Upload guideline
- Upload notifications
  - Upload started and in progress
  - Upload failed to complete (stopped, either by user or error such as loss of connection)
  - Upload complete

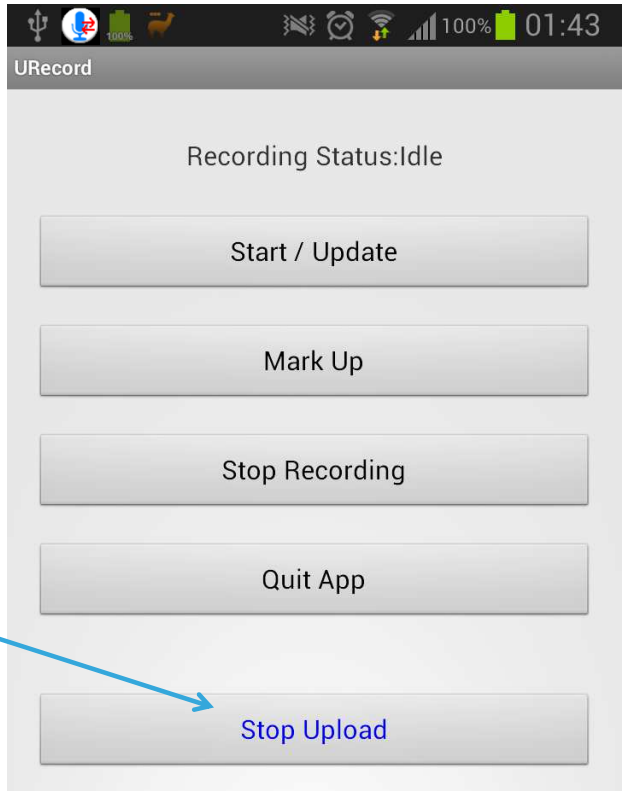


life.augmented

Step #6: Upload data

17

- Once you've finished a recording, you can upload data to our server
- From Home view, click Upload
- Upload guideline
- Upload notifications
- When upload is started, the Upload button now displays a Stop Upload which allows to stop it



life.augmented

Step #7: Uninstall URecord

18

- In Settings, Application Manager, find Urecord Application
  - Click it and select Uninstall
  - This will remove app and configuration files
- HOWEVER, data files are still stored on memory
  - To remove them, find Urecord directory in the File Manager (also called My Files)
  - Remove all files from the directory

What we expect you to do

19

- Make continuous recordings long of a few hours (4-5 hours max)
- Visit different scenes during a single recording
- Scenes we are interested in:
  - Home, Office, different transportation means (car, bus, tram, train, bicycle), Restaurant/Pub, Walking in Street, Shop
- Live your life as usual and use your smartphone as usual
- How to make annotations
  - Annotations should be made at scene transitions, annotate the new scene
  - E.g. when leaving Home and going outdoors in the street, annotate the scene Street
  - E.g. when entering a building (home, workplace, shop, etc) or a mean of transportation

life.augmented

Scenario Examples

20

- A workday, start recording in the morning, leave home and take a transportation mean for going to work, keep recording at work, stop recording at lunch
- A workday, start recording during the afternoon while at work, keep recording while returning home in the evening, stop recording at home
- A weekend, start recording at home, then leave home and take a mean of transport, do some shopping, go back home and stop recording
- Start a recording in a place, then leave it, enter a restaurant for lunch/dinner, leave the restaurant and when back home, stop recording

life.augmented

Anonymity of Data

- Sensitive data sources
  - Audio (risk to record people's conversations)
  - Location providers (risk to record user's home location coordinates)
  - Character strings (risk to record application names, phone numbers, electronic device identifiers such as Bluetooth ones or Wifi ones
- Scramble data or keep features
  - Audio: keep features only (ZCR, DFT magnitude coefficients)
  - Location providers: shift coordinates
  - Character strings: hash strings
- Every data collected earlier than 48 hours before new record is deleted
- Wifi transfer: data is stored in anonymous repositories
  - One per user, randomly created name